

Повышаем живучесть Raft в реальных условиях



Сергей
Останевич








Меню

- **Raft overview**
  - Термины: Journal, Term, LSN
  - Выборы
  - Гарантии
  - Ожидания != Реальность
- Raft / Tarantool: особенности
- Надстройки Raft
  - Pre-Vote
  - Split-Vote detection
  - Fencing





Raft overview: Термины

- **Raft** – алгоритм достижения консенсуса в распределенных системах
- Raft достигает консенсуса за счет **выбора единого лидера**, который может изменять состояние системы
- Лидер выбирается в рамках **term** – пронумерованного отрезка времени
- Лидер может добавлять записи в **журнал (log)**
- Каждая запись в журнале пронумерована **log index/log sequence number (LSN)**





Raft overview: Выборы

Follower

timeout,
start elections

Candidate

quorum of votes
received, promote

Leader

Node with higher term or
leader in current term discovered

Node with higher term discovered,
resign leadership





Raft overview: Выборы

Follower

timeout,
start elections

Candidate

quorum of votes
received, promote

Leader

Node with higher term or
leader in current term discovered

Node with higher term discovered,
resign leadership





Raft overview: Выборы

Follower

timeout,
start elections

Candidate

quorum of votes
received, promote

Leader

Node with higher term or
leader in current term discovered

Node with higher term discovered,
resign leadership





Raft overview: Выборы

Follower → (timeout, start elections) → **Candidate** → (quorum of votes received, promote) → Leader

Candidate → Follower: **Node with higher term or leader in current term discovered**

Leader → Follower: Node with higher term discovered, resign leadership





Raft overview: Выборы

Follower

timeout,
start elections

Candidate

**quorum of votes
received, promote**

Leader

Node with higher term or
leader in current term discovered

Node with higher term discovered,
resign leadership





Raft overview: Выборы

Follower

timeout,
start elections

Candidate

quorum of votes
received, promote

Leader

Node with higher term or
leader in current term discovered

Node with higher term discovered,
resign leadership





Raft overview: Выборы

Follower

timeout,
start elections

Candidate

quorum of votes
received, promote

Leader

Node with higher term or
leader in current term discovered

**Node with higher term discovered,
resign leadership**





Raft overview: Гарантии

- **Election Safety**: один лидер в каждом term
- **Leader Append-Only**: лидер только добавляет записи без модификации или удаления записей
- **Log Matching**: записи с идентичными индексом и term в двух журналах гарантируют их совпадение вплоть до этих записей
- **Leader Completeness**: запись в журнал в каком-либо term будет присутствовать в логах всех последующих лидеров
- **State Machine Safety**: запись, примененная с определенным индексом на одном сервере, будет такой же и на другом сервере





Raft overview: Гарантии

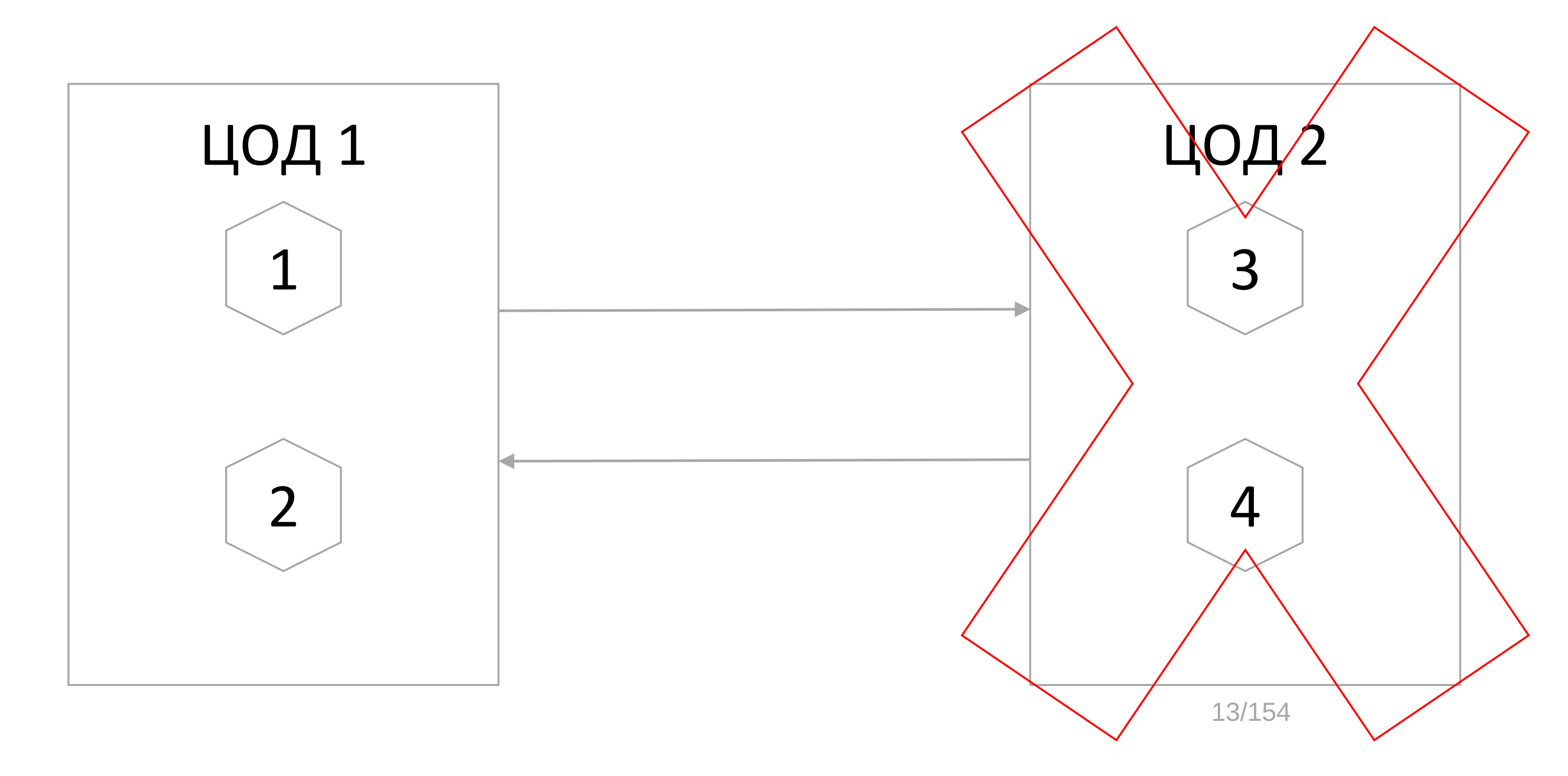
- **Рано или поздно** лидер найдётся (скорее всего)
- **В рамках одного term** есть не более одного лидера





Raft overview: Гарантии

- **Рано или поздно** лидер найдётся (скорее всего)
- **В рамках одного term** есть не более одного лидера







Raft overview: Гарантии

- **Рано или поздно** лидер найдётся (скорее всего)
- **В рамках одного term** есть не более одного лидера

При ограничениях

- В кластере есть нода, имеющая необходимое количество соединений, чтобы выиграть выборы

$$T_{ping} << T_{election} << T_{accident}$$





Raft overview: Ожидания != Реальность

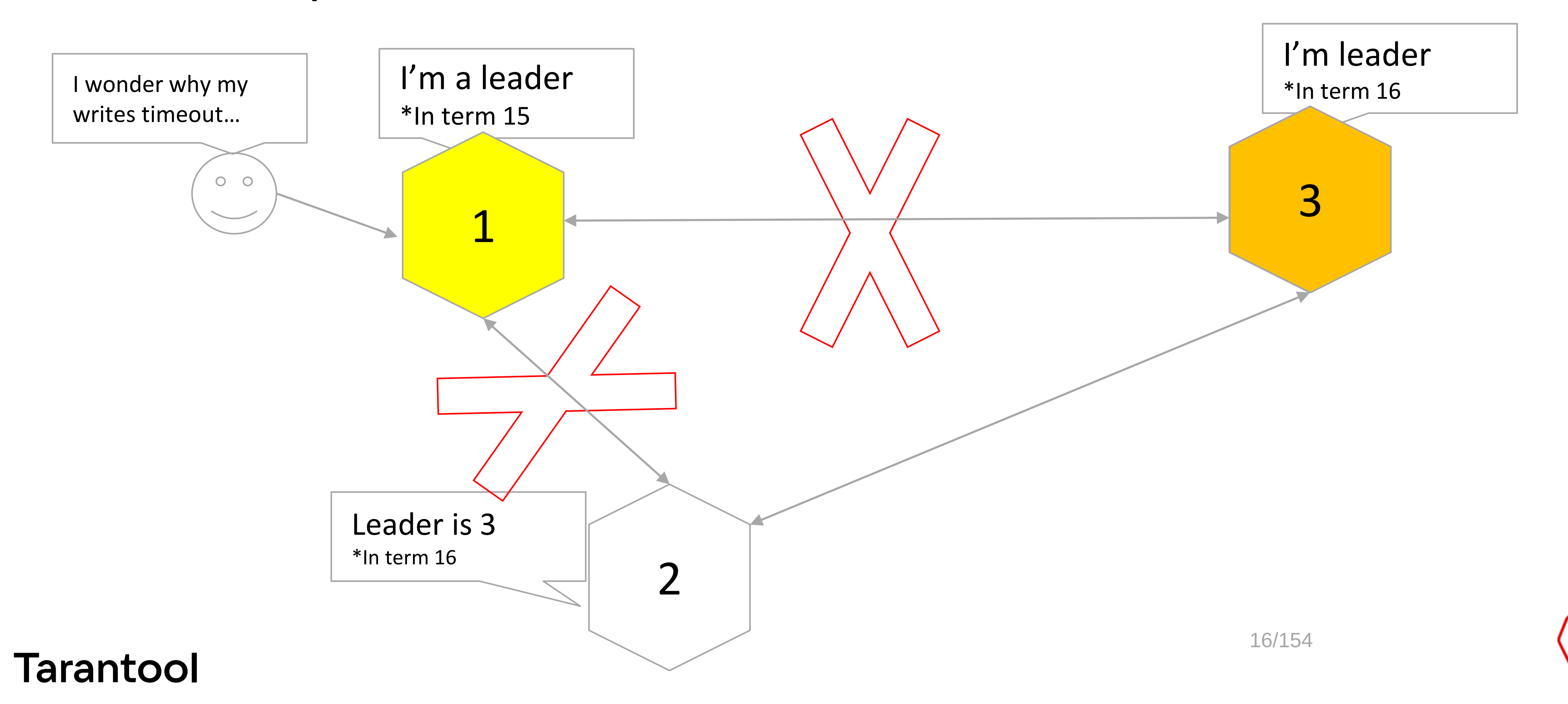
- Рано или поздно != быстро
- Term != время





Raft overview: Ожидания != Реальность

- Рано или поздно != быстро
- Term != время



Tarantool

HL HighLoad++ 2022

Raft overview: Ожидания != Реальность

Ожидаем:

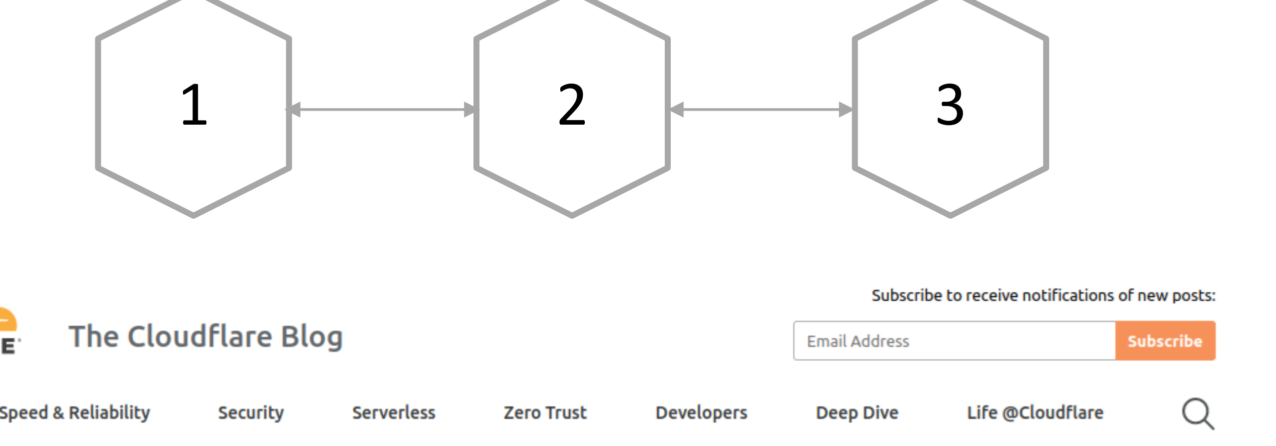
- Найти лидера **быстро**
- Не более одного лидера в кластере в **конкретный момент времени**
- Возвращение/подключение нод в кластер не вызывает его простоя





Raft overview: Ожидания != Реальность

В процессе тестирования Raft в Tarantool мы нашли проблему в такой конфигурации:



И почти сразу нашли, что мы не одиноки

A Byzantine failure in the real world

27.11.2020

On November 2, 2020, Cloudflare had an incident that impacted the availability of the API and dashboard for six hours and 33 minutes. During this incident, the





Вторая перемена блюд

- ✓ Raft overview
  - Journal, Term, LSN
  - Выборы
  - Гарантии
  - Ожидания != Реальность
- **Raft / Tarantool: особенности**
- Надстройки Raft
  - Pre-Vote
  - Split-Vote detection
  - Fencing





Tarantool: Мультиплексированный журнал

1	2	3	4	5	6

log index = 6

1:1	**1:2**	**2:1**	**1:3**	**1:4**	**1:5**

vclock = {1 : 5, 2: 1}

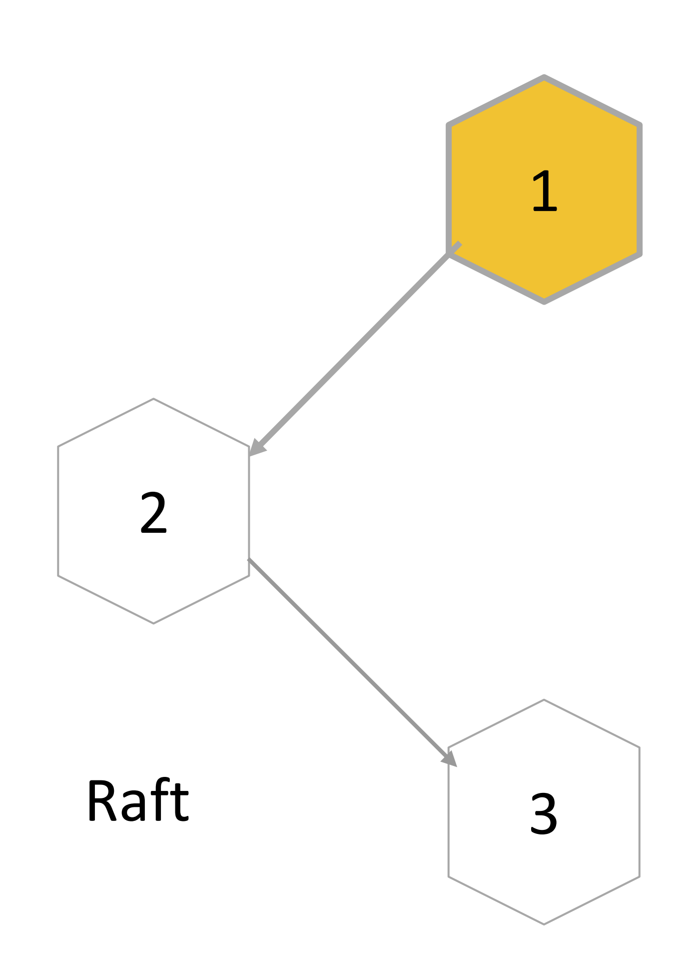
Term

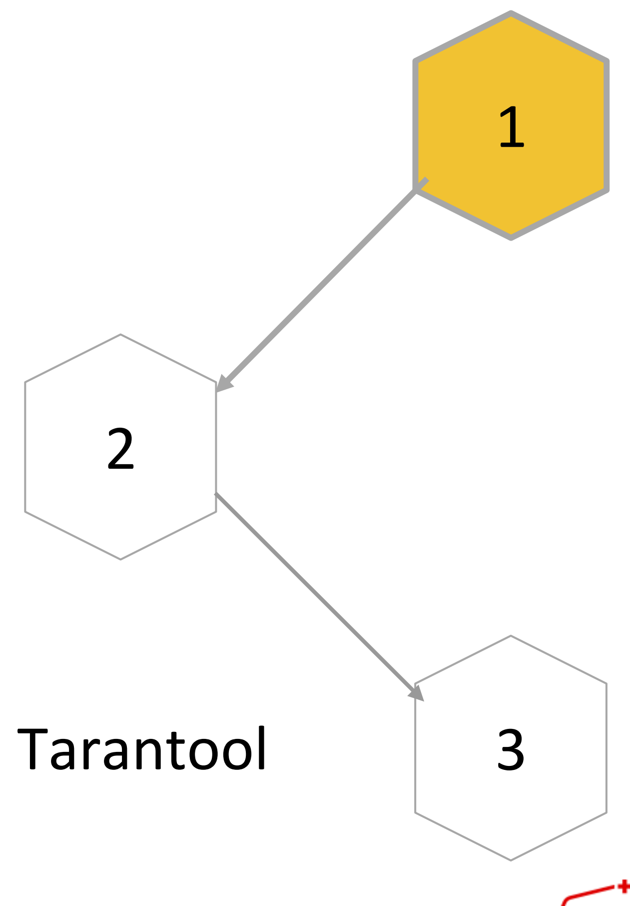
Node	1	2	3
1			
2			





Tarantool: Рассылка данных от всех ко всем

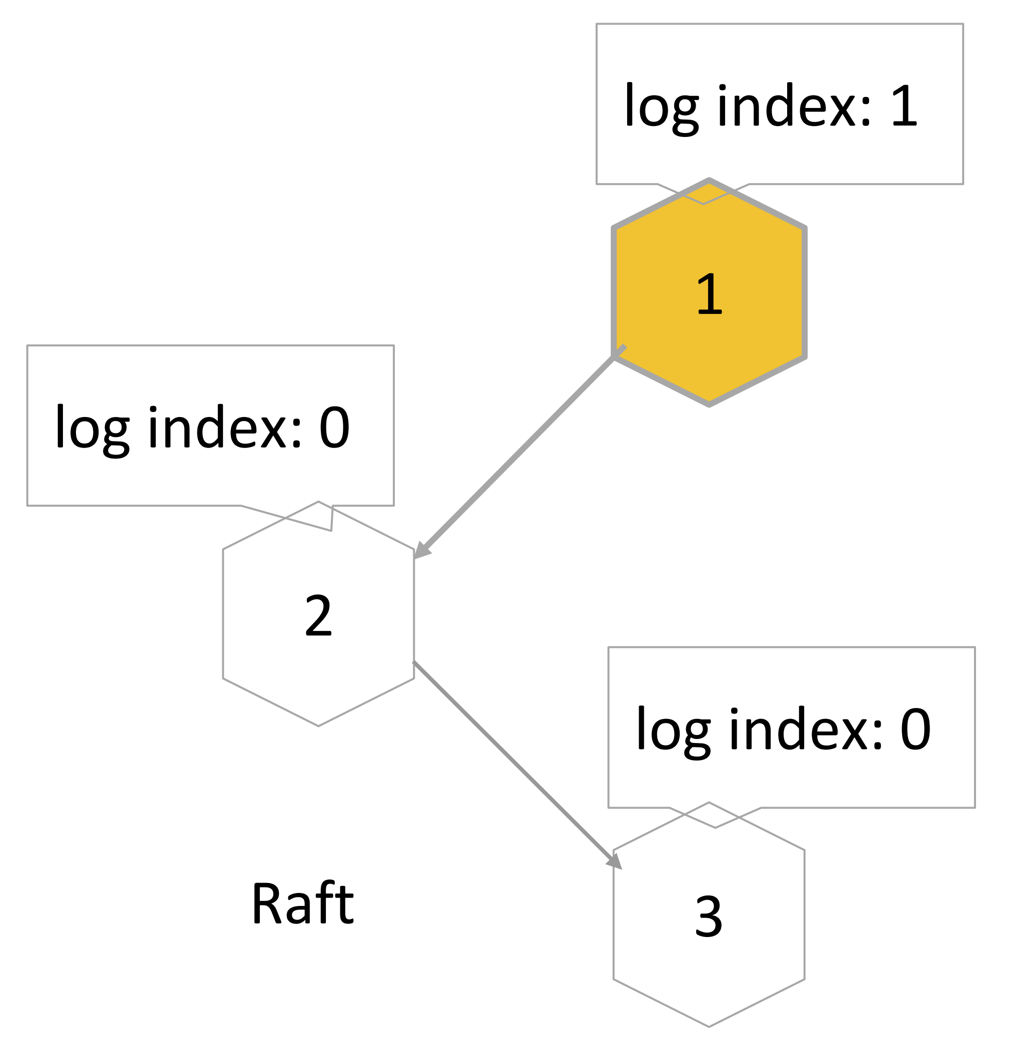








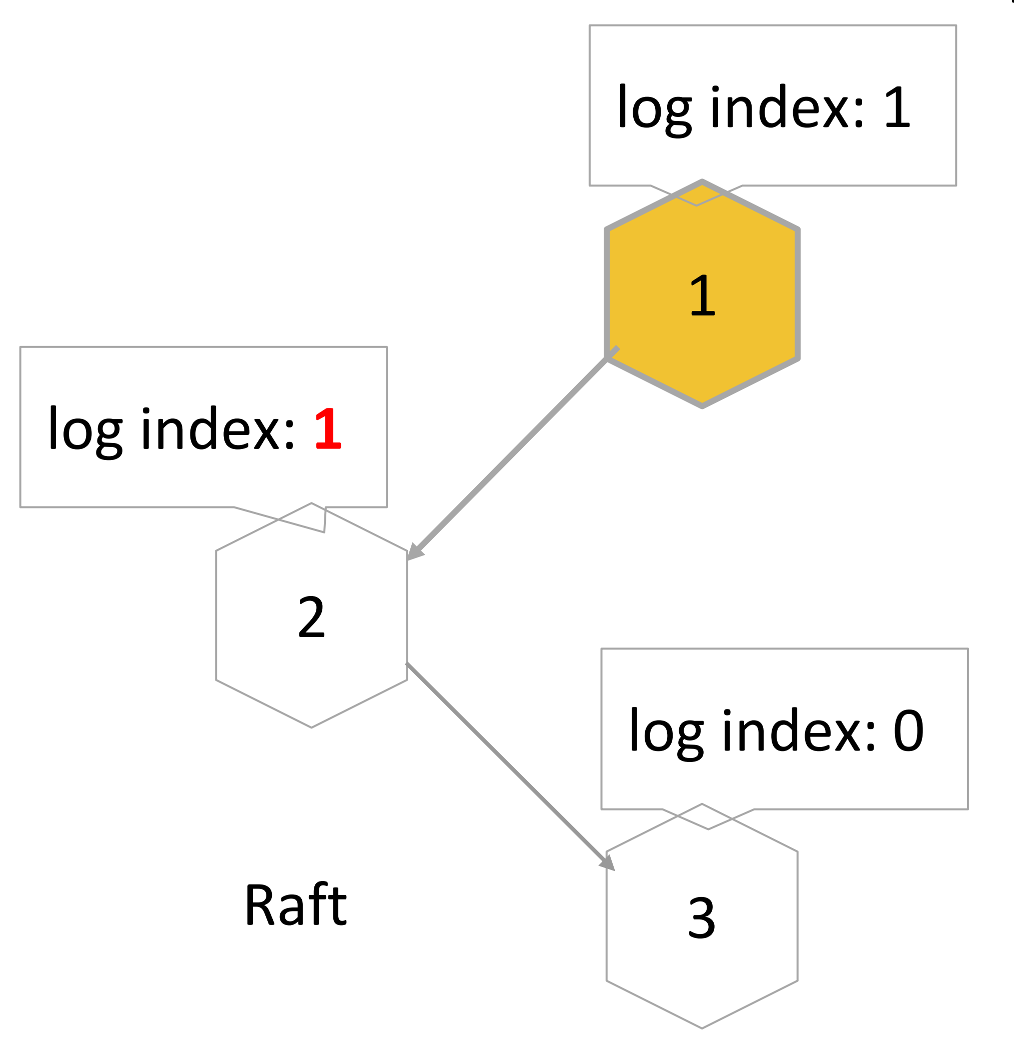
Tarantool: Рассылка данных от всех ко всем







Tarantool: Рассылка данных от всех ко всем



vclock: {1:1, 2:0, 3:0}

1

vclock: {**1:1**, 2:0, 3:0}

2

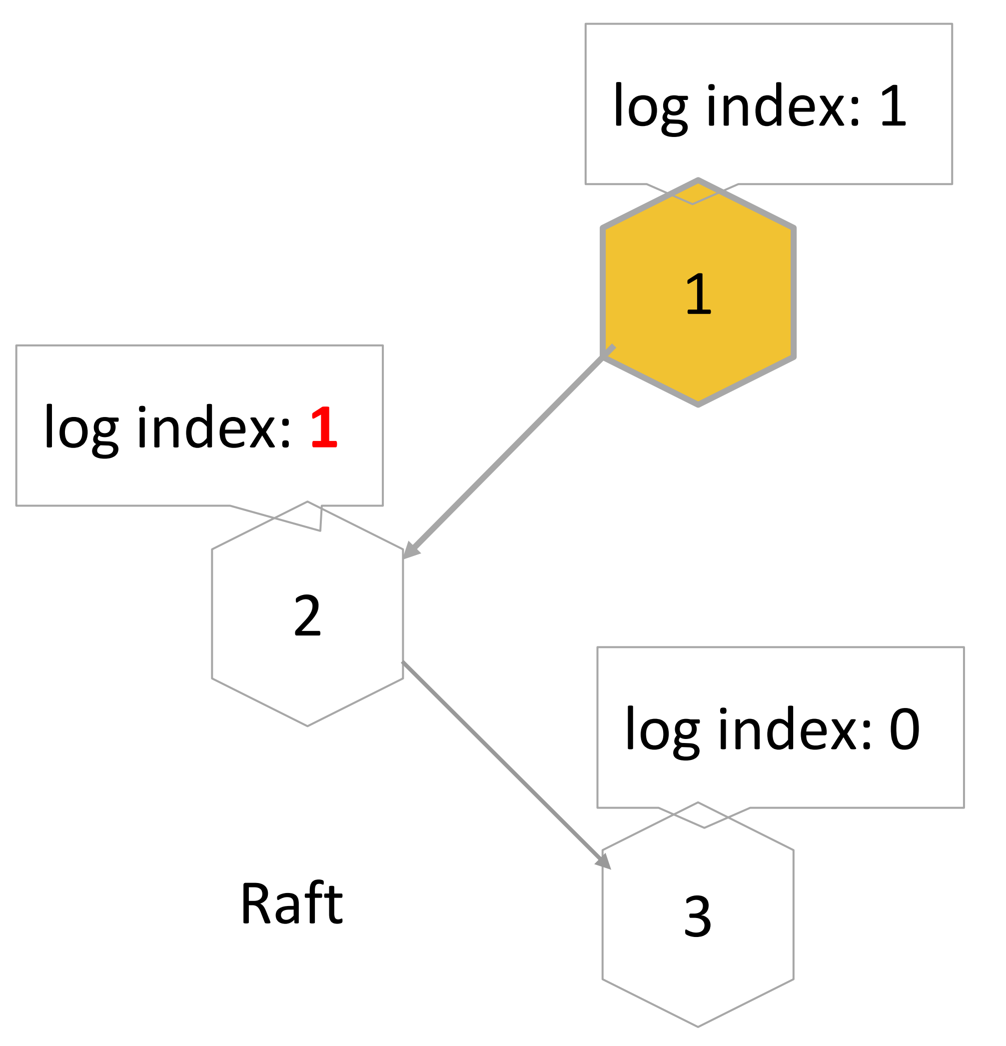
vclock: {1:0, 2:0, 3:0}

Tarantool

3



Tarantool: Рассылка данных от всех ко всем



vclock: {1:1, 2:0, 3:0}

1

vclock: {**1:1**, 2:0, 3:0}

2

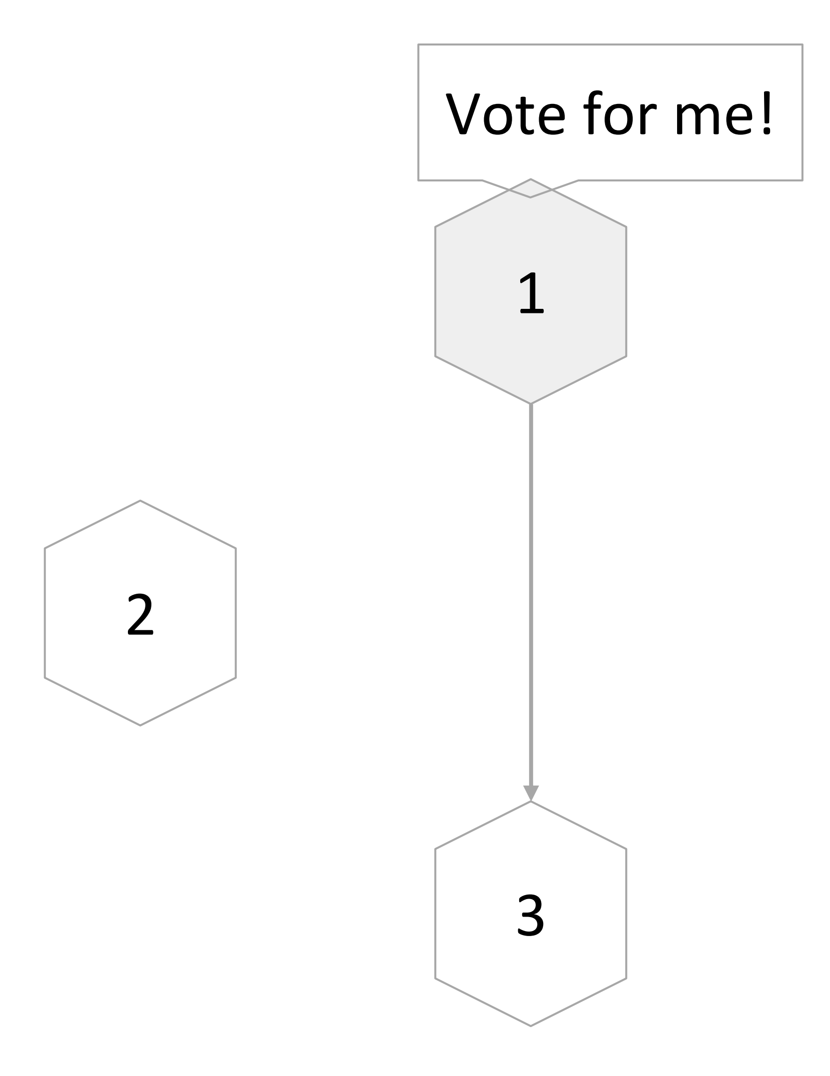
vclock: {**1:1**, 2:0, 3:0}

Tarantool

3



Tarantool: Броадкаст результатов голосования







Tarantool: Броадкаст результатов голосования

Vote for me!

1

Ok, 3 voted for 1

2

I voted for 1

3

Сообщения существующих типов можно дополнять без нарушения обратной совместимости





Меню

- ✓ Raft overview
  - Термины: Journal, Term, LSN
  - Выборы
  - Гарантии
  - Ожидания != Реальность
- ✓ Raft / Tarantool: особенности
- **Надстройки Raft**
  - **Pre-Vote**
  - Split-Vote detection
  - Fencing





Pre-Vote: Проблема №1

Недоступность кластера на запись после восстановления связности





Проблема №1: Пример

Term 3

1

I'm the leader

2

1 is leader

Term 3

3

1 is leader

Term 3





Проблема №1: Пример

Term 3

1

2

Term 3

1 is leader

I'm the leader

Where's leader?

3

Term 3





Проблема №1: Пример

Term 3

1

I'm the leader

2

Term 3

1 is leader

3

Term **4**

Starting elections





Проблема №1: Пример

1 is leader

Term 3

1

2

Term 3

I'm the leader

Where are my votes?

3

Term 4





Проблема №1: Пример

Term 3

1

2

Term 3

1 is leader

I'm the leader

Starting elections

3

Term 5





Проблема №1: Пример

Term 3

1

I'm the leader

2

1 is leader

Term 3

3

Where are my votes?

Term **5**





Проблема №1: Пример

Term 3

1

Term 3

2

Where's leader?

New term! Resign

Where are my votes?

3

Term 5

Кластер недоступен для записи





Проблема №1: Хотелось бы

Term 3

1

I’m the leader

1 is leader

2

Term 3

Where’s leader?

3

Term 3





Проблема №1: Хотелось бы

Term 3

1

I’m the leader

2

1 is leader

Term 3

Waiting for quorum

3

Term 3





Проблема №1: Хотелось бы

Term 3

1

I'm the leader

1 is leader

2

Term 3

1 is leader

3

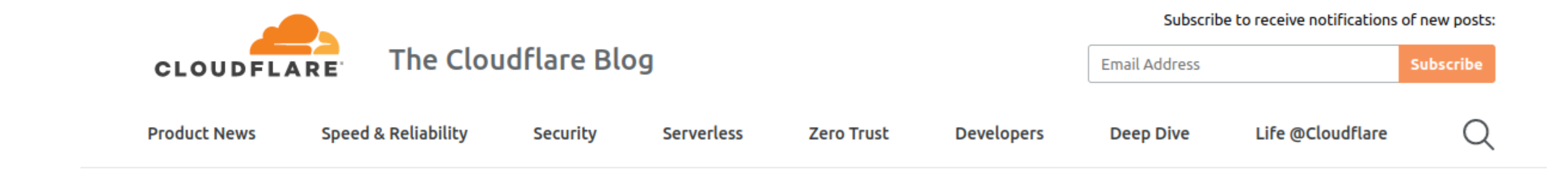
Term 3





Pre-Vote: Проблема №2

Сервер, который не видит лидера, будет постоянно его прерывать



A Byzantine failure in the real world

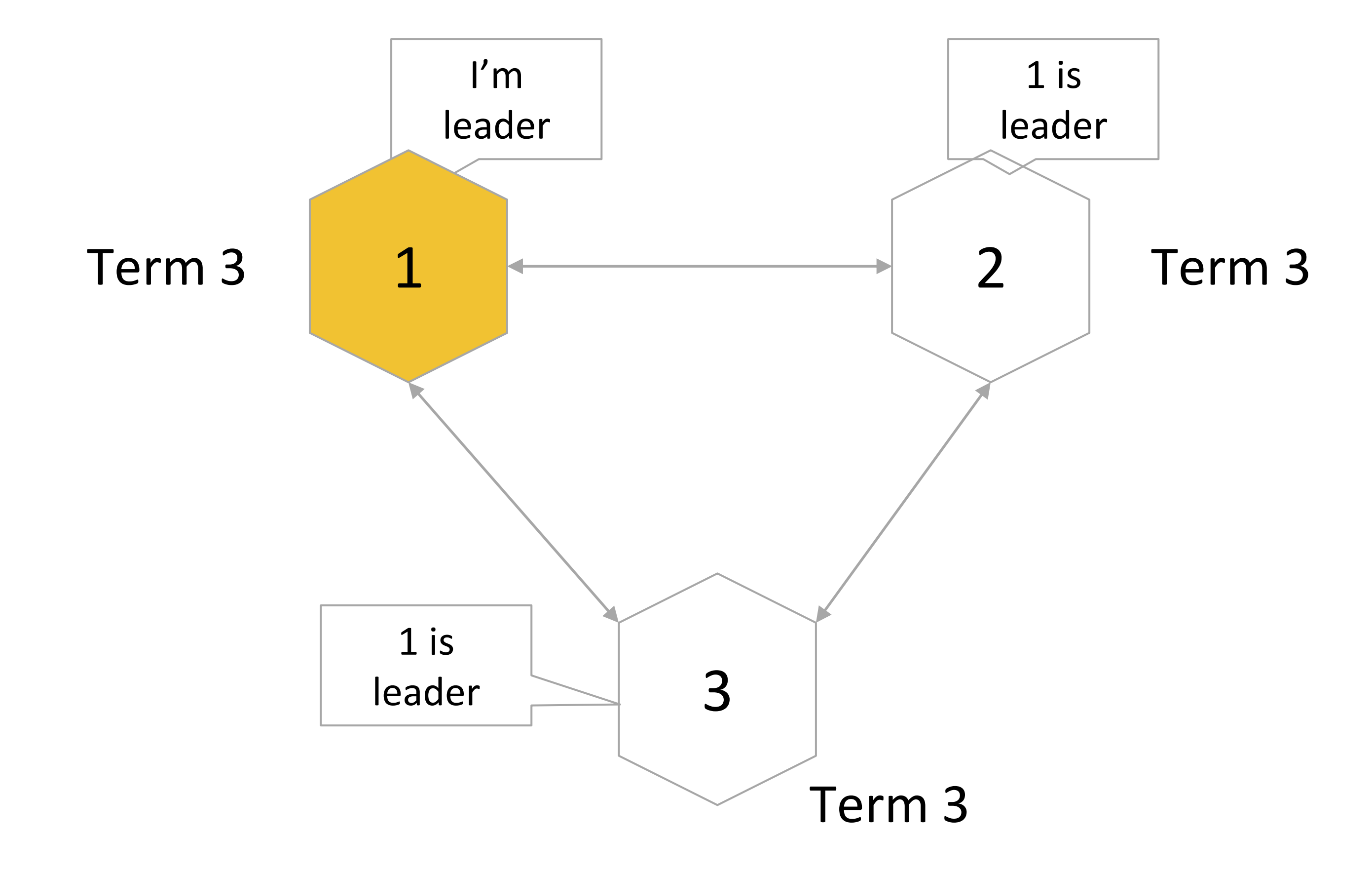
27.11.2020

On November 2, 2020, Cloudflare had an incident that impacted the availability of the API and dashboard for six hours and 33 minutes. During this incident, the





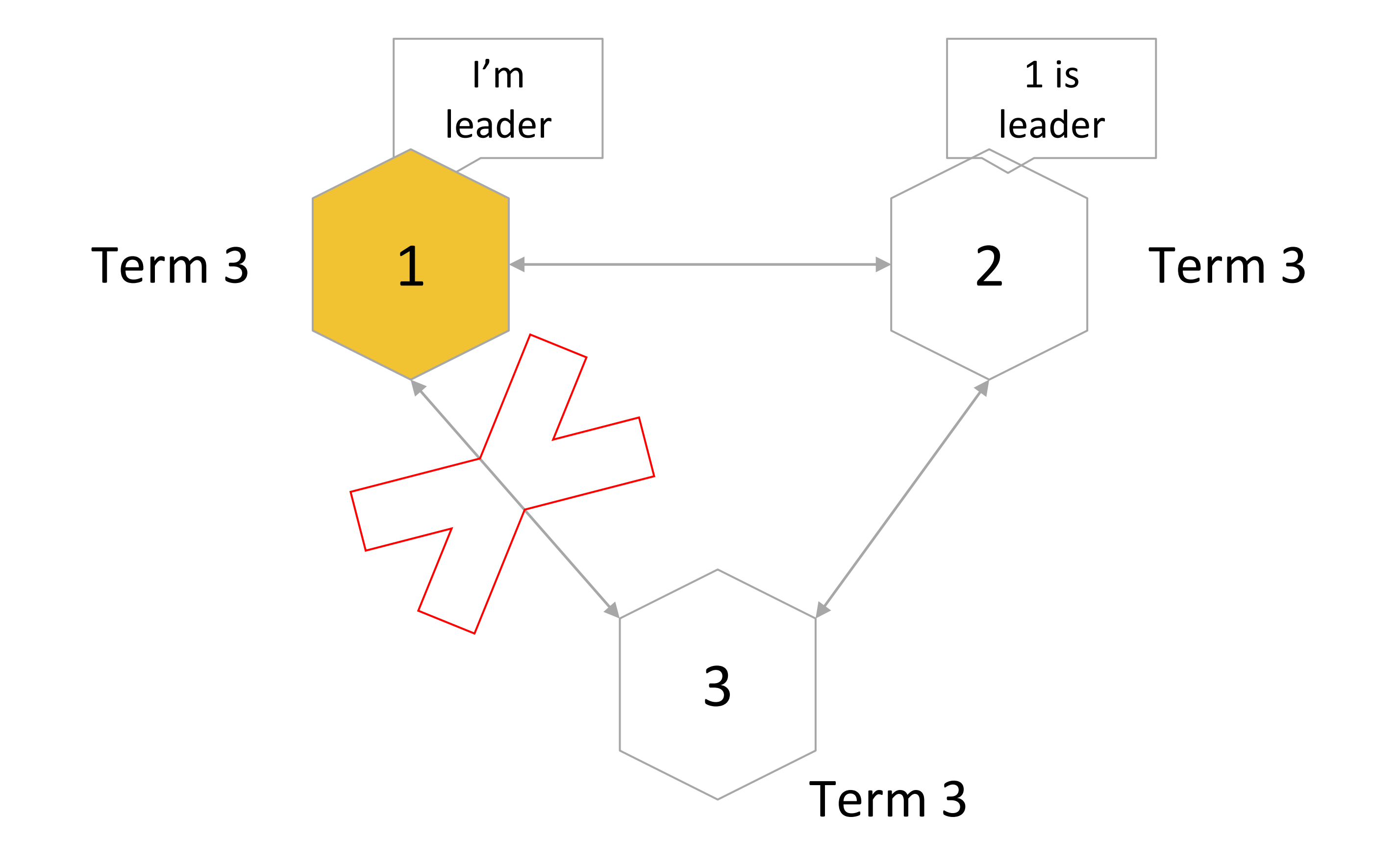
Проблема №2: Пример







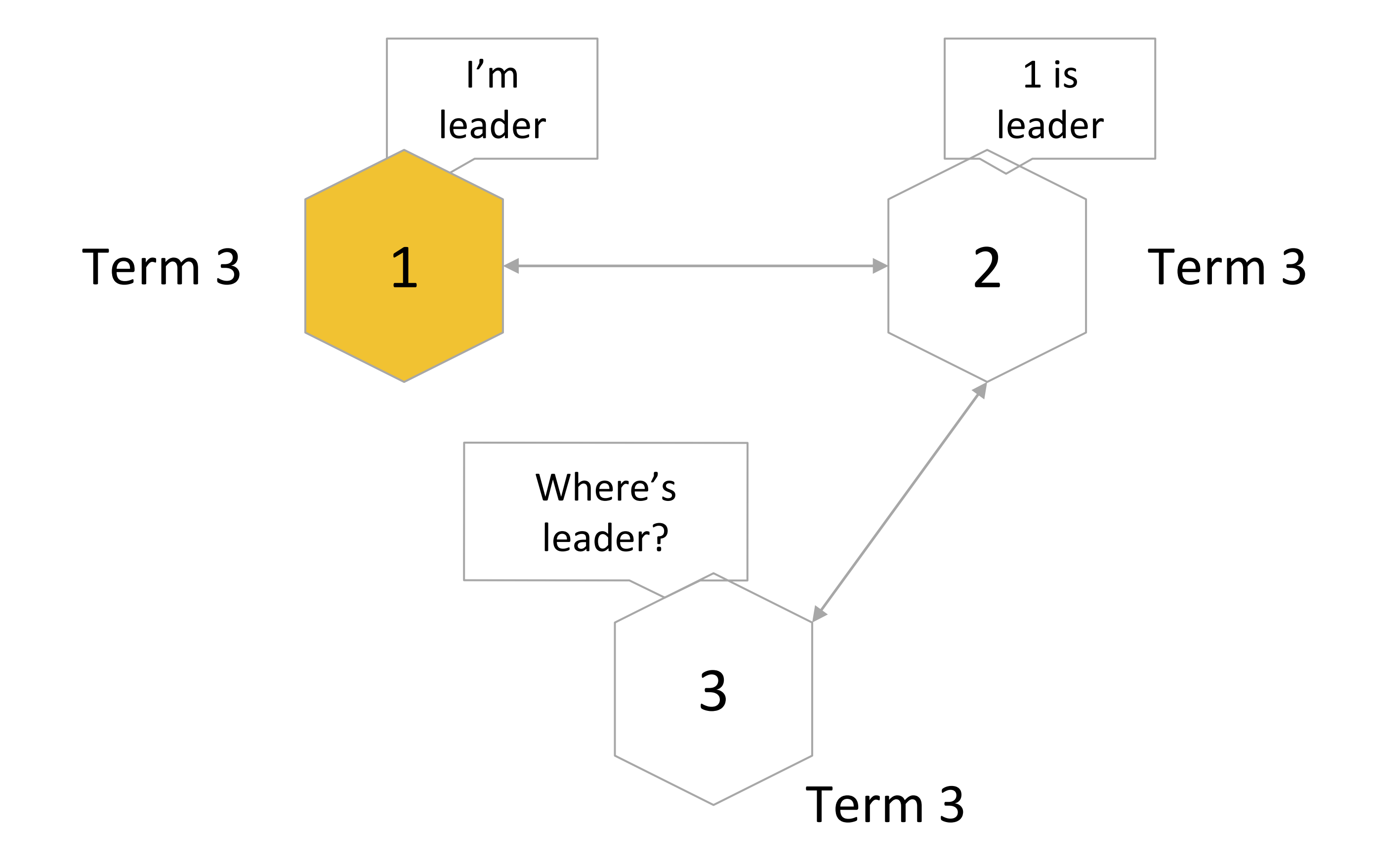
Проблема №2: Пример







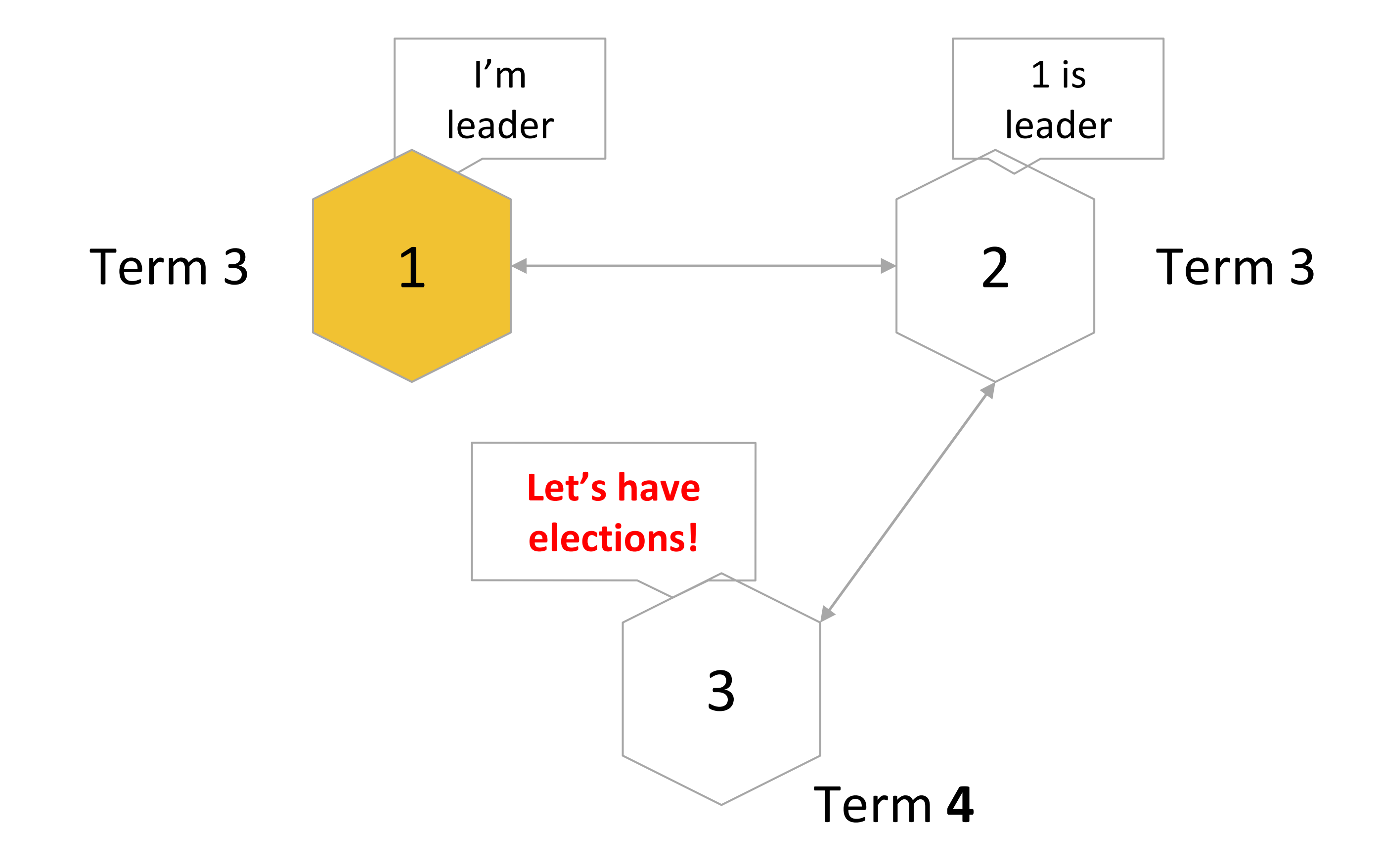
Проблема №2: Пример







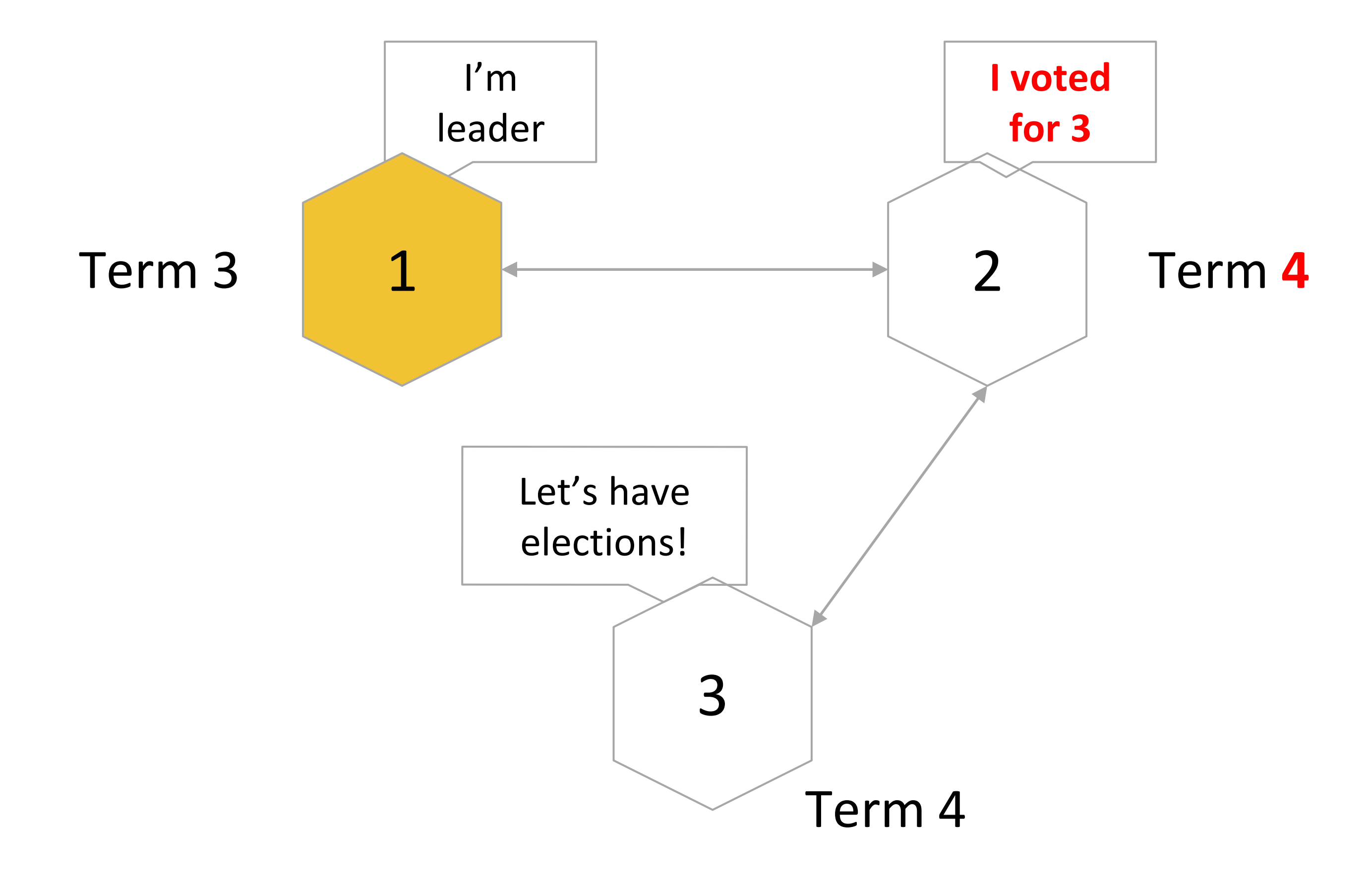
Проблема №2: Пример







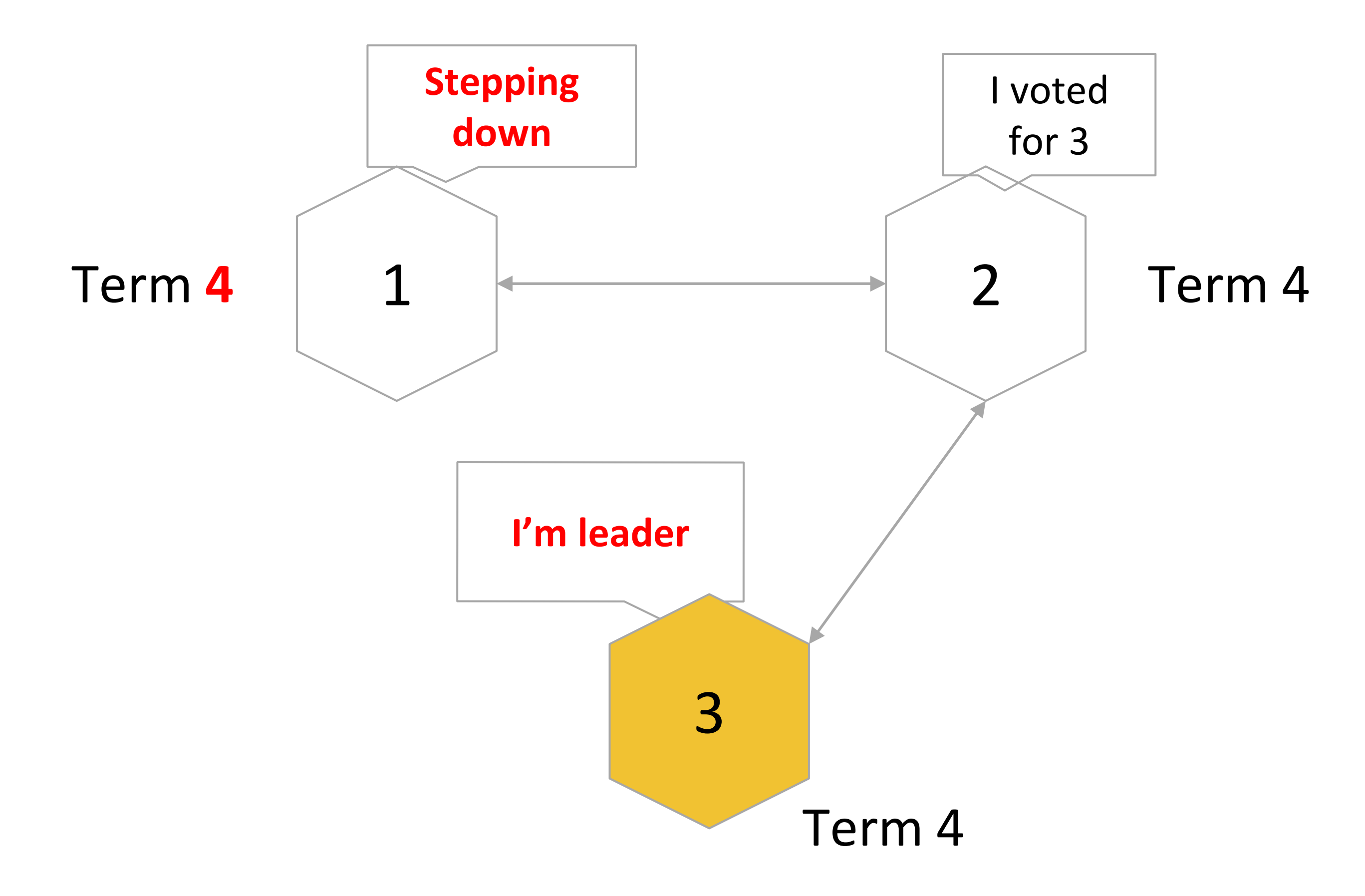
Проблема №2: Пример



Tarantool



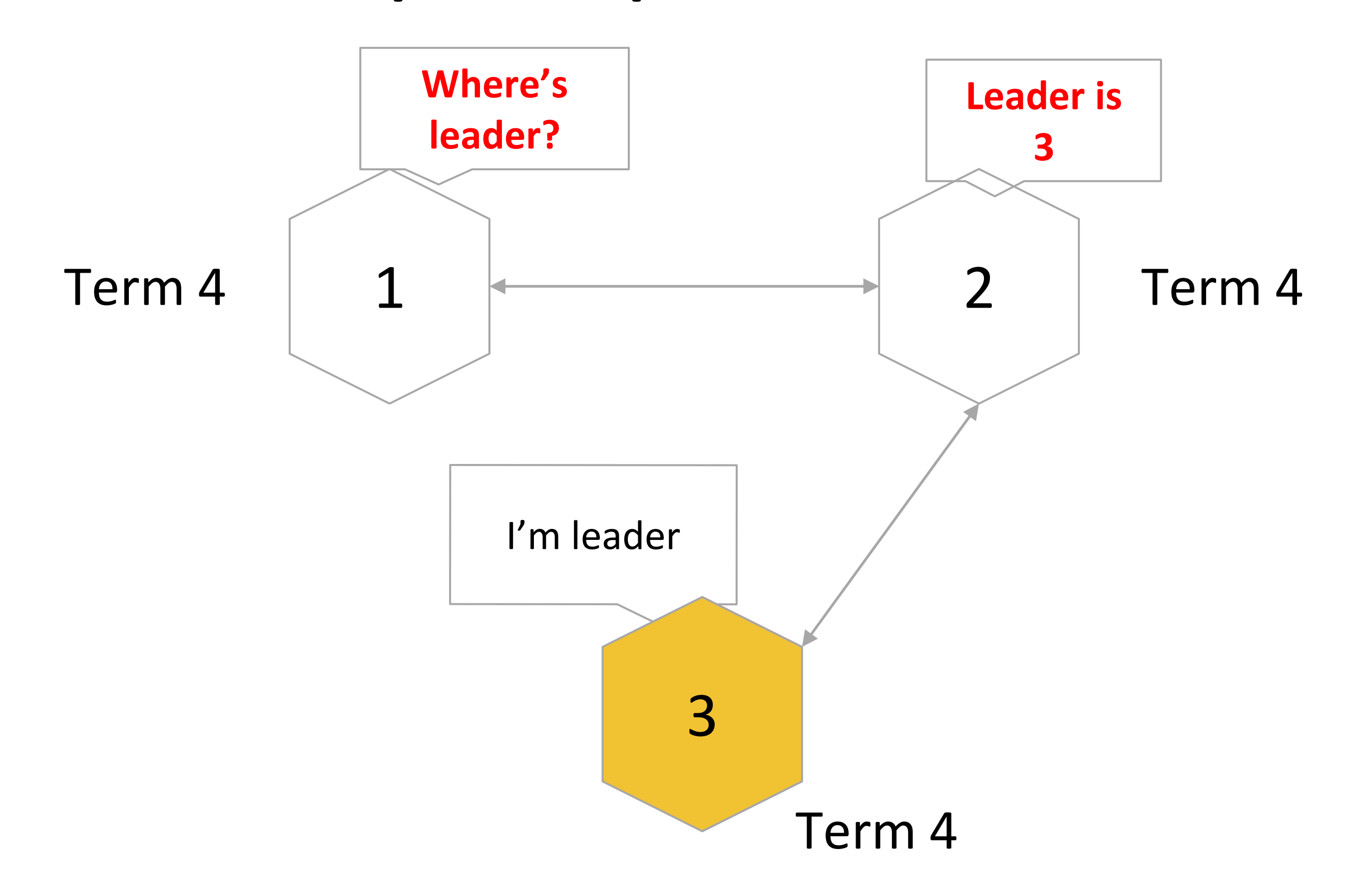
Проблема №2: Пример







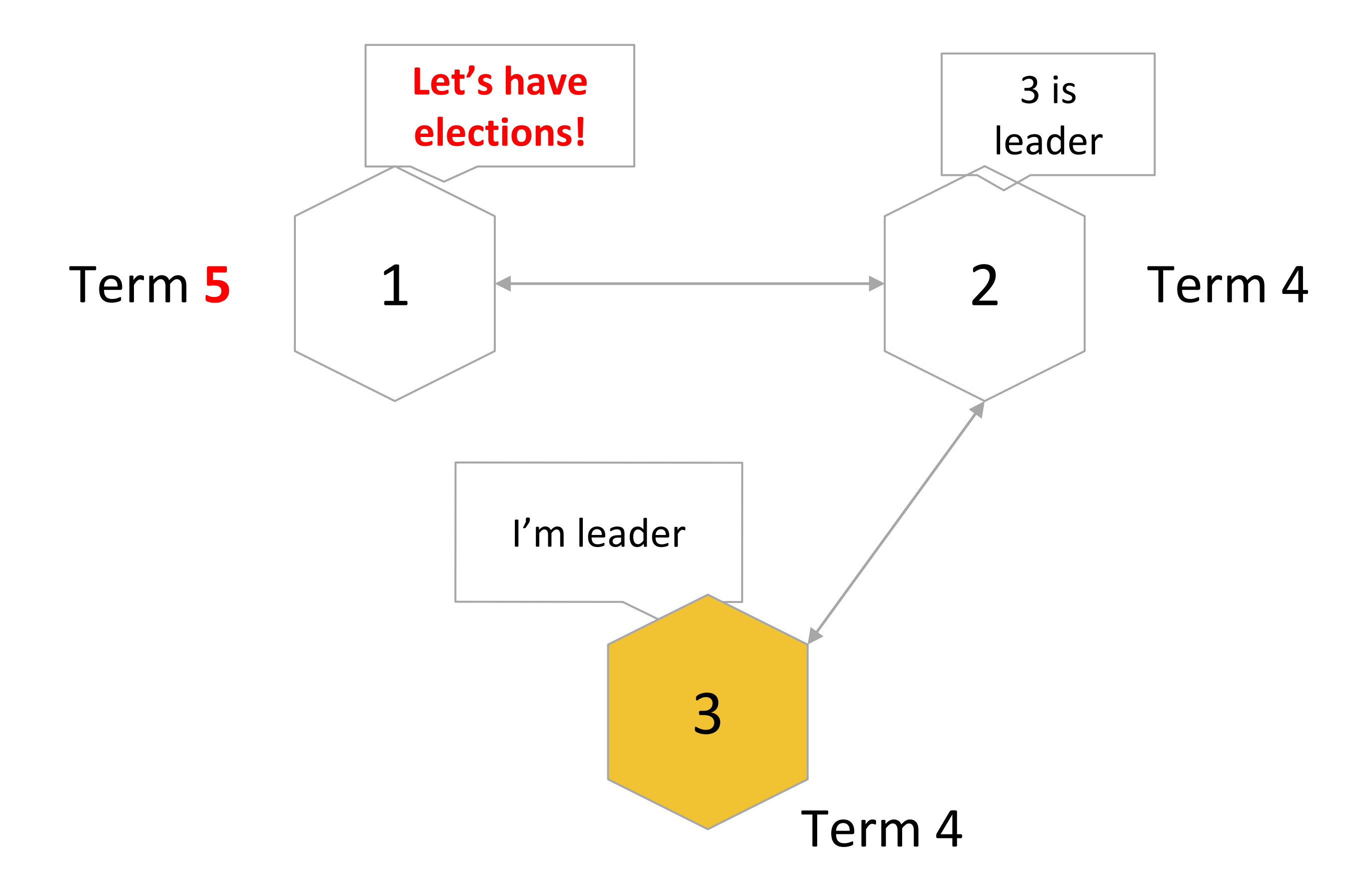
Проблема №2: Пример







Проблема №2: Пример



Tarantool



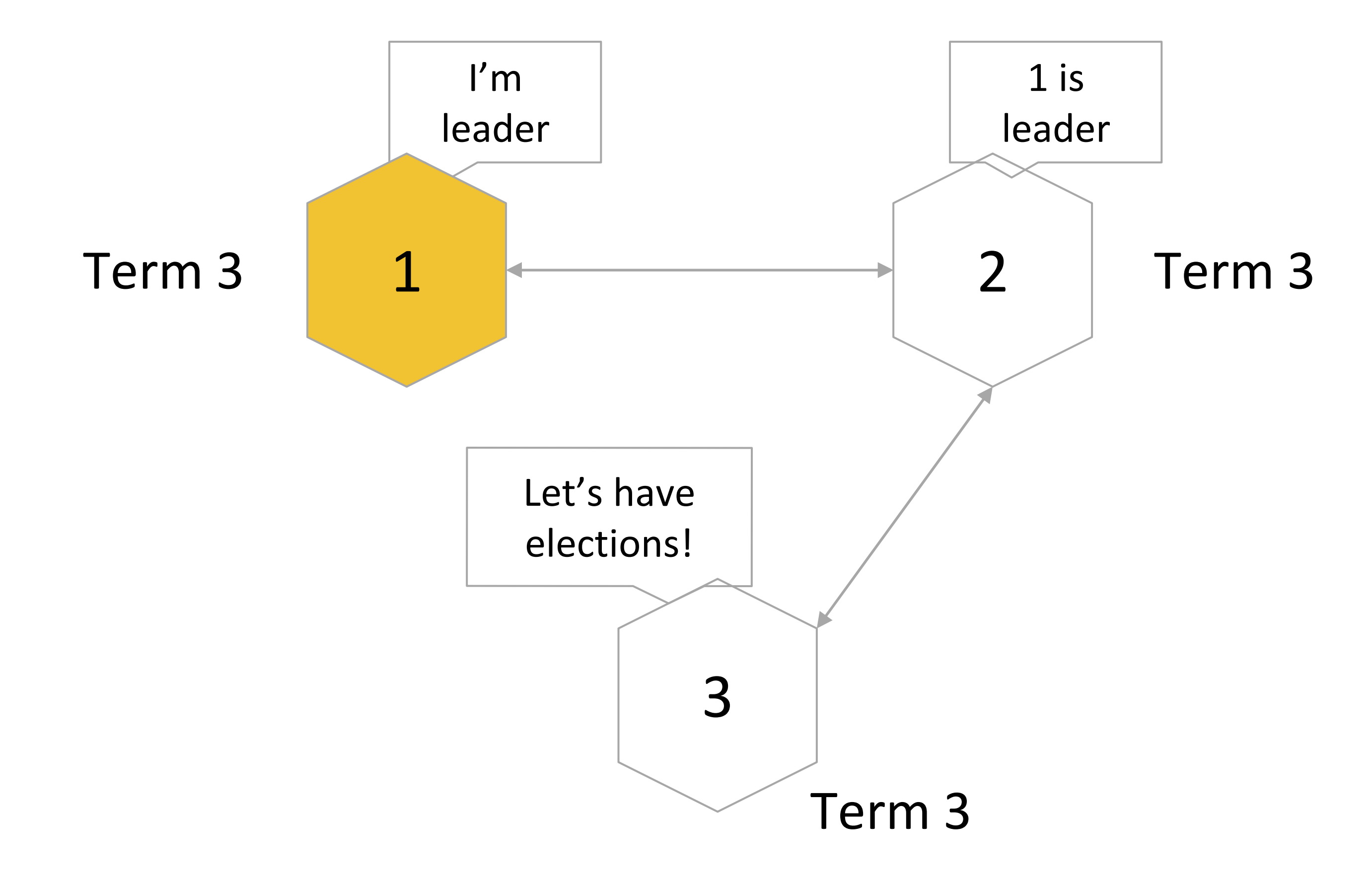
Проблема №2: Пример

Кластер постоянно перевыбирает лидера





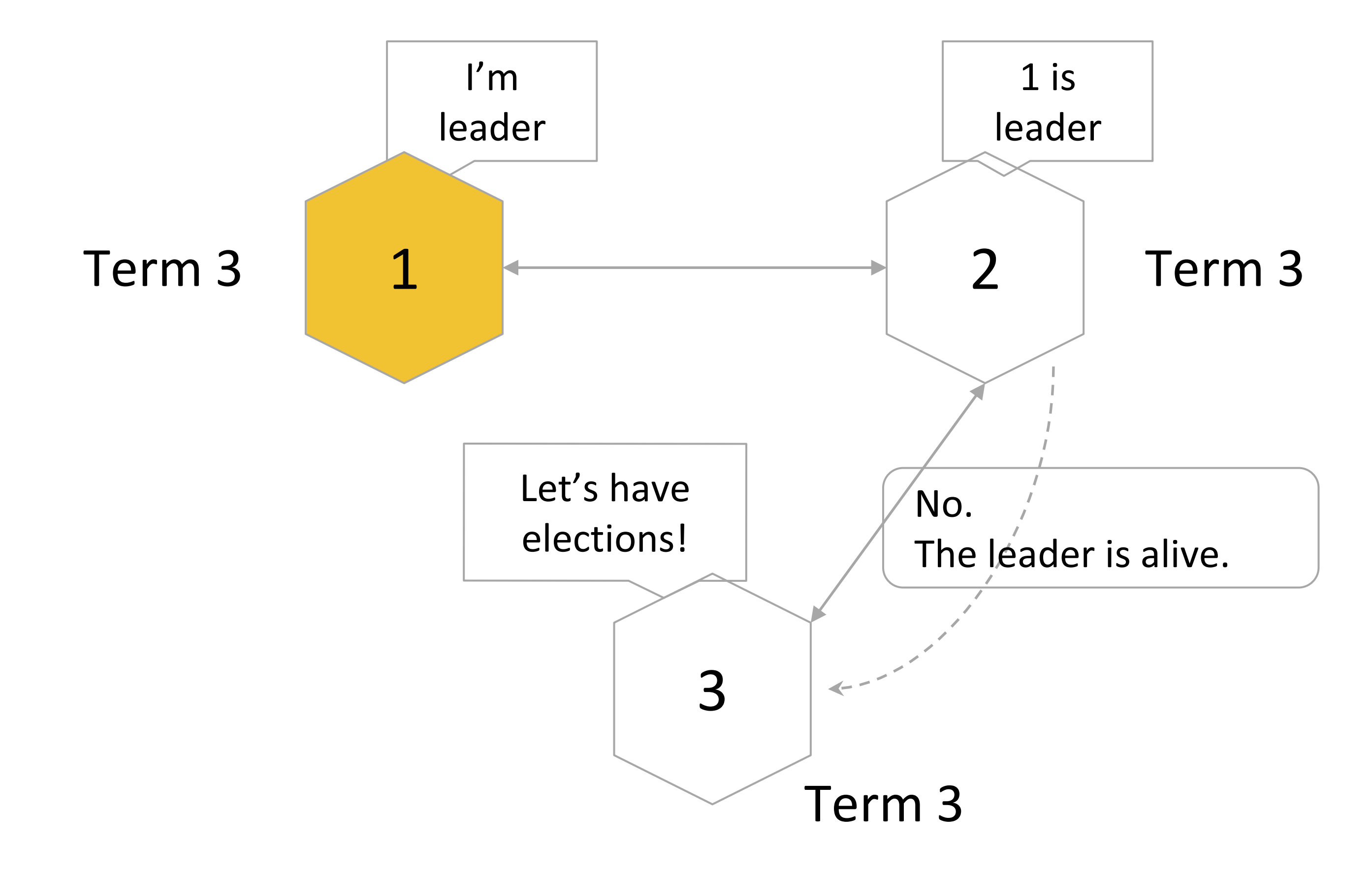
Проблема №2: Как хочется







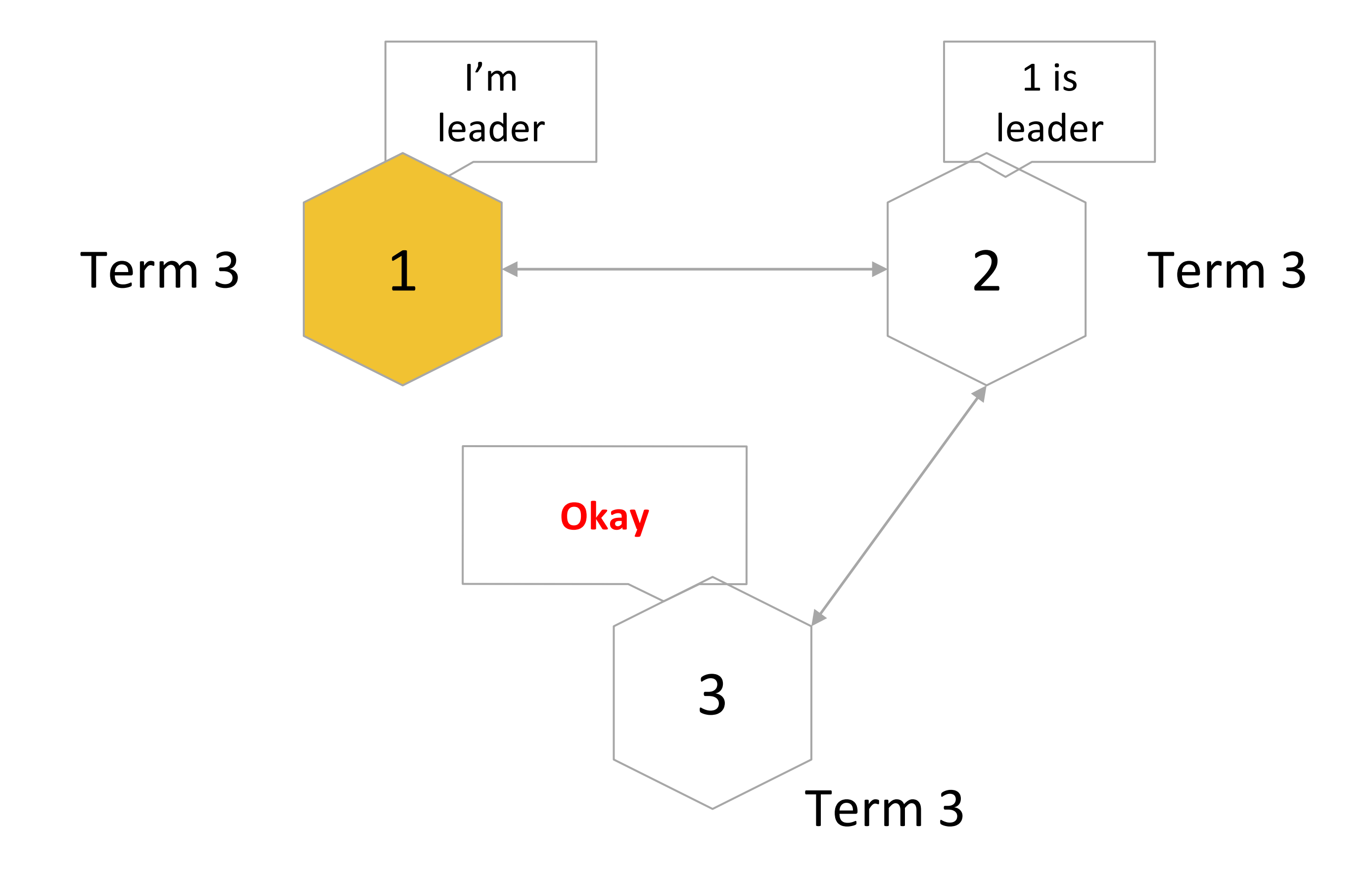
Проблема №2: Как хочется







Проблема №2: Как хочется



Tarantool



Проблема №2: Как хочется

Итак, хотим, чтобы действующего лидера не мог прервать никто





Pre-Vote: Варианты решения

1. **Предварительные выборы**
   - Перед началом выборов спрашиваем, проголосуют ли за нас

2. Игнорирование выборов голосующими, которые видят лидера
3. Pre-Vote на основе метаинформации





Pre-Vote: Варианты решения

1. **Предварительные выборы**
   - Перед началом выборов спрашиваем, проголосуют ли за нас
   - Ответ положительный, если голосующий сам не видит лидера, и если проголосовал бы за кандидата в обычных выборах

2. Игнорирование выборов голосующими, которые видят лидера
3. Pre-Vote на основе метаинформации





Pre-Vote: Варианты решения

1. **Предварительные выборы**
    - Перед началом выборов спрашиваем, проголосуют ли за нас
    - Ответ положительный, если голосующий сам не видит лидера, и если проголосовал бы за кандидата в обычных выборах
    - После получения кворума положительных ответов начинаем выборы
2. Игнорирование выборов голосующими, которые видят лидера
3. Pre-Vote на основе метаинформации





Предварительные выборы: Пример №1

1 is leader

Term 3

1

2

Term 3

I'm the leader

Where's leader?

3

Term 3





Предварительные выборы: Пример №1

1 is leader

Term 3

1

2

Term 3

I'm the leader

Waiting for quorum

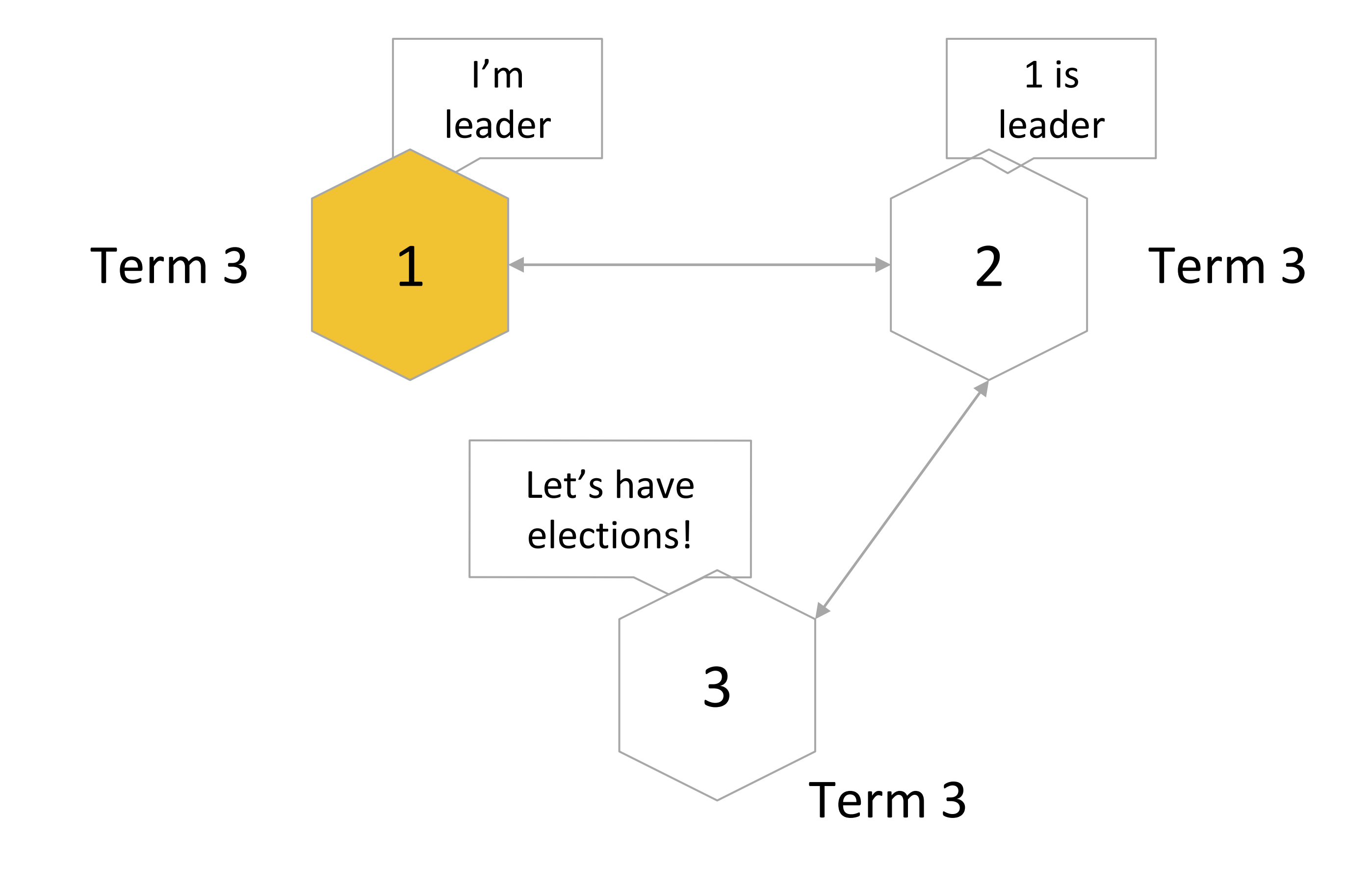
3

Term 3





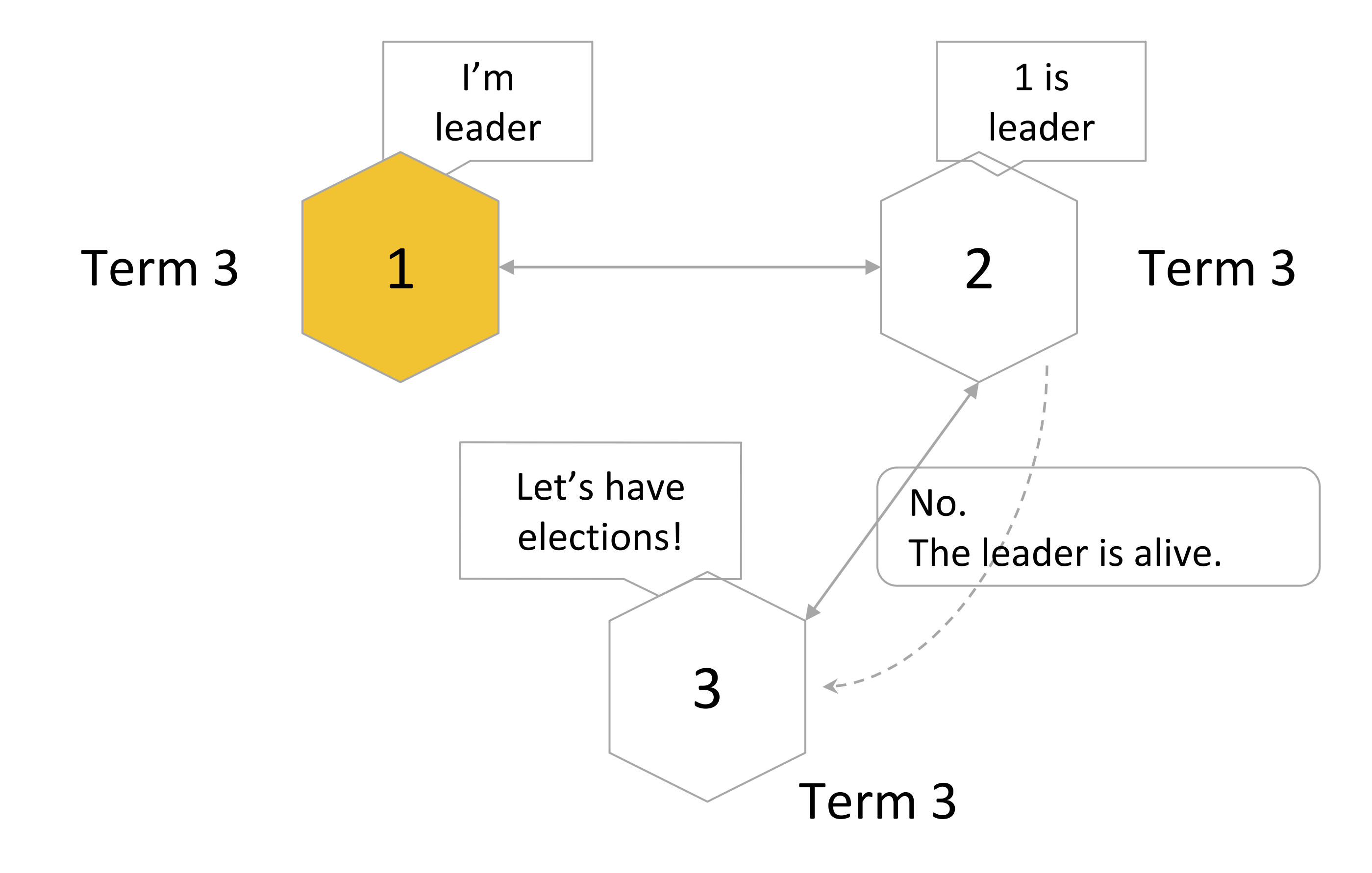
Предварительные выборы: Пример №2







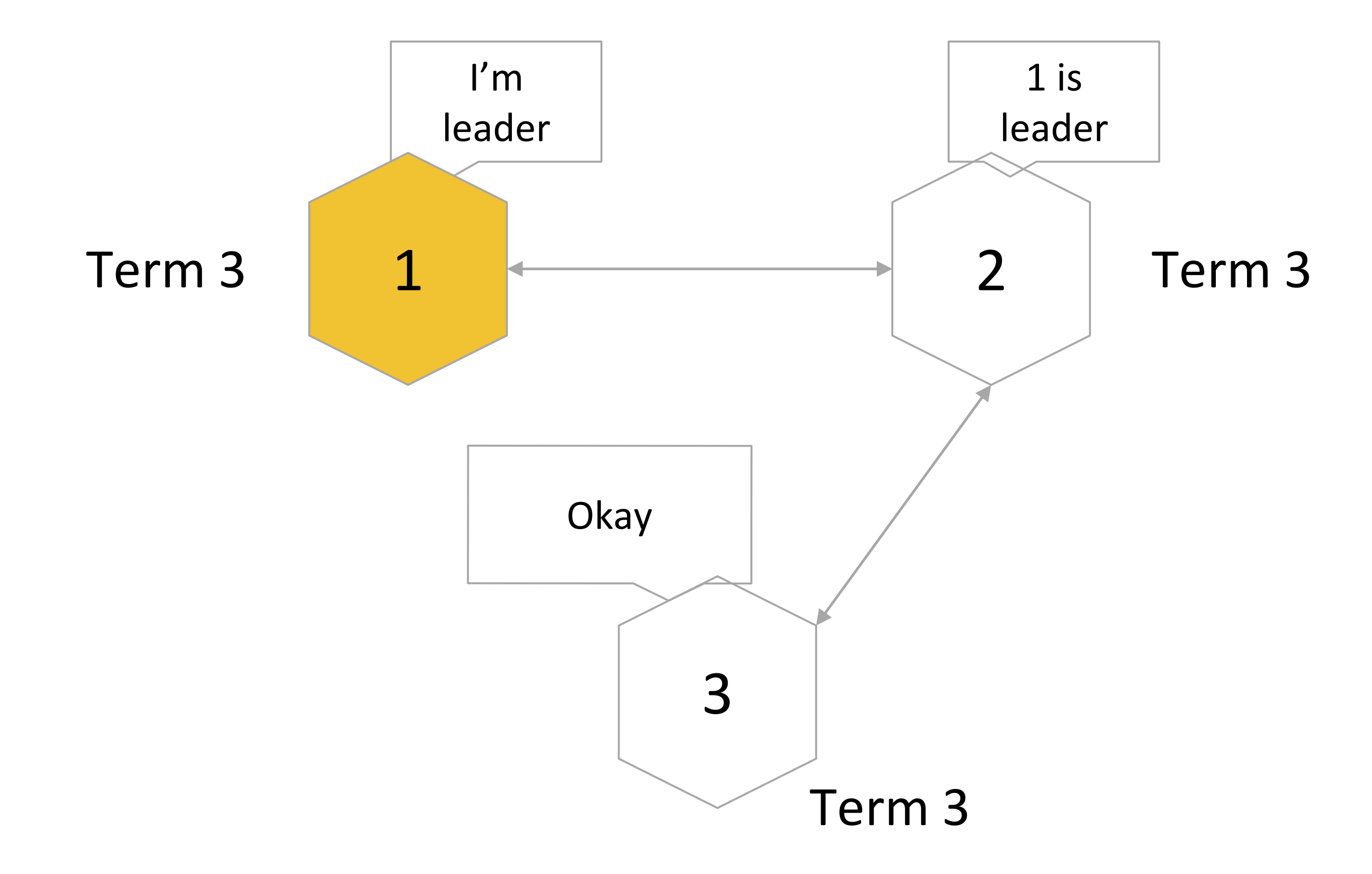
Предварительные выборы: Пример №2







Предварительные выборы: Пример №2







Pre-Vote: Варианты решения

1. **Предварительные выборы**
   - Перед началом выборов спрашиваем, проголосуют ли за нас
   - Ответ положительный, если голосующий сам не видит лидера и если проголосовал бы за кандидата в обычных выборах
   - После получения кворума положительных ответов начинаем выборы
2. Игнорирование выборов голосующими, которые видят лидера
3. Pre-Vote на основе метаинформации

Недостаток – требуется новый тип сообщений, ломающий обратную совместимость с существующими инсталляциями





Pre-Vote: Варианты решения

1. Предварительные выборы
2. **Игнорирование выборов голосующими, которые видят лидера**

3. Pre-Vote на основе метаинформации





Pre-Vote: Варианты решения

1. Предварительные выборы
2. **Игнорирование выборов голосующими, которые видят лидера**
   - Сервер, видящий лидера напрямую, игнорирует любые запросы на голос от соседей
3. Pre-Vote на основе метаинформации





Игнорирование: Пример №1

Term 3

1

I'm the leader

1 is leader

2

Term 3

Where's leader?

3

Term 3





Игнорирование: Пример №1

1 is leader

Term 3

1

2

Term 3

I'm the leader

Let's have elections!

3

Term **4**





Игнорирование: Пример №1

Term 3

1

I'm the leader

1 is leader

2

Term 3

Where are my votes?

3

Term **4**





Игнорирование: Пример №1

1 is leader

Term 3

1

2

Term 3

Stepping down

Where are my votes?

3

Term 4





Pre-Vote: Варианты решения

1. Предварительные выборы
2. **Игнорирование выборов голосующими, которые видят лидера**
   - Сервер, видящий лидера напрямую, игнорирует любые запросы на голос от соседей
3. Pre-Vote на основе метаинформации

Недостаток – решение неполное!





Pre-Vote: Варианты решения

1. Предварительные выборы
2. Игнорирование выборов голосующими, которые видят лидера
3. **Pre-Vote на основе метаинформации**
    - Каждый сервер знает состояние Raft соседей за счет широковещательной рассылки





Pre-Vote: Варианты решения

1. Предварительные выборы
2. Игнорирование выборов голосующими, которые видят лидера
3. **Pre-Vote на основе метаинформации**
    - Каждый сервер знает состояние Raft соседей за счет широковещательной рассылки
    - Добавим в сообщение о состоянии информацию о том, виден ли лидер





Pre-Vote: Варианты решения

1. Предварительные выборы
2. Игнорирование выборов голосующими, которые видят лидера
3. **Pre-Vote на основе метаинформации**
   - Каждый сервер знает состояние Raft соседей за счет широковещательной рассылки
   - Добавим в сообщение о состоянии информацию о том, виден ли лидер
   - Если кто-то видит лидера в текущем терме – не начинаем выборы





Pre-Vote: Пример №1

Term 3

1

2

Term 3

1 is leader

I'm the leader

Waiting for quorum

3

Term 3





Pre-Vote: Пример №1

Term 3

1

I'm the leader

2

Term 3

1 is leader

3

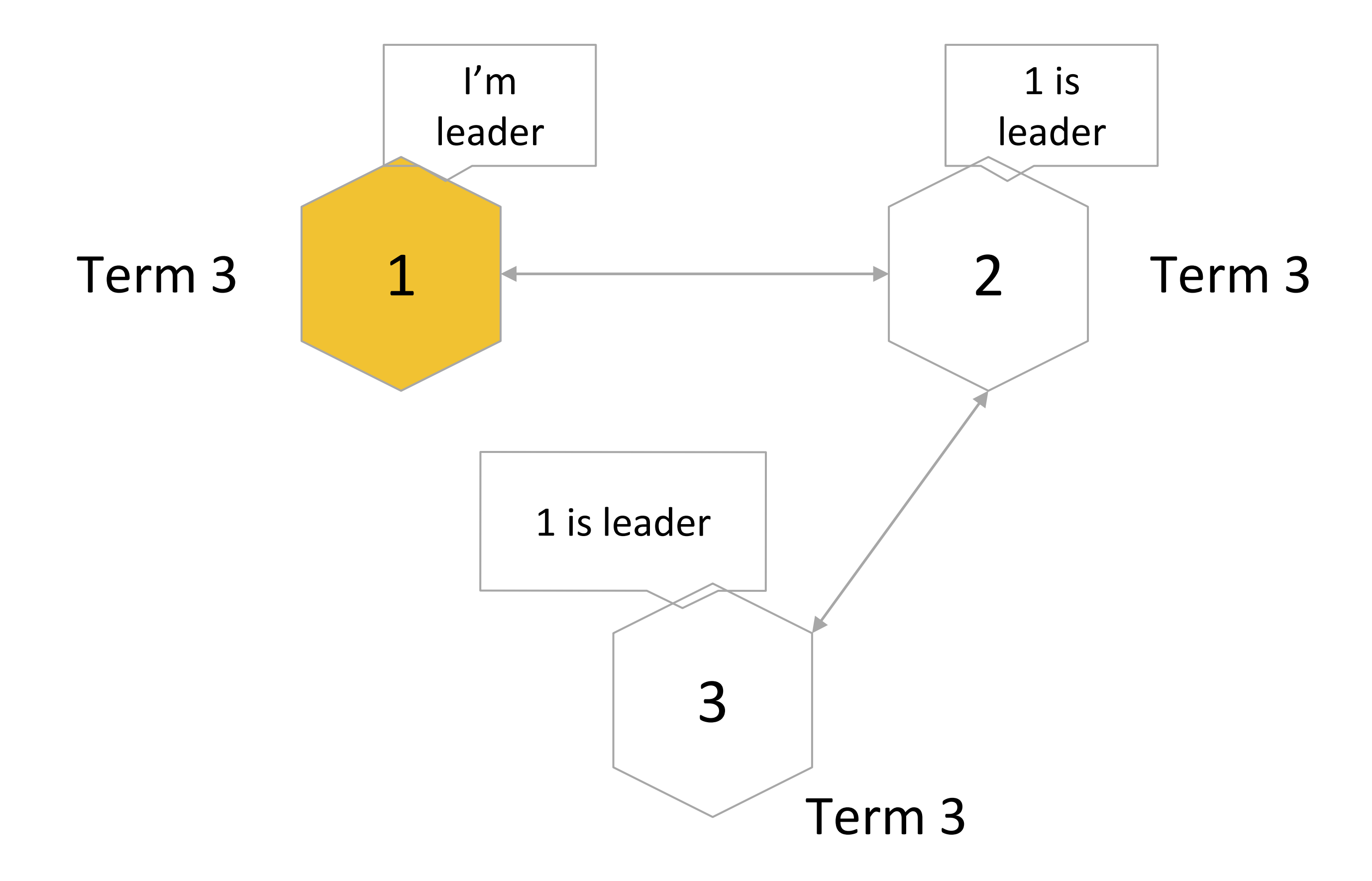
Term 3

1 is leader





Pre-Vote: Пример №2







Pre-Vote: Варианты решения

1. Предварительные выборы
2. Игнорирование выборов голосующими, которые видят лидера
3. **Pre-Vote на основе метаинформации**
    - Каждый сервер знает состояние Raft соседей за счет широковещательной рассылки
    - Добавим в сообщение о состоянии информацию о том, виден ли лидер
    - Если кто-то видит лидера в текущем терме - не начинаем выборы

Решение как с предварительными выборами, но обратно совместимое





Меню

- ✓ Raft overview
  - Термины: Journal, Term, LSN
  - Выборы
  - Гарантии
  - Ожидания != Реальность
- ✓ Raft / Tarantool: особенности
- Надстройки Raft
  - ✓ Pre-Vote
  - **Split-Vote detection**
  - Fencing





Split-Vote: Проблема

Долгая недоступность кластера на запись после потери лидера





Split-Vote: Проблема

Долгая недоступность кластера на запись после потери лидера

Каждый раунд выборов, в результате которого не выбран лидер –
≥ 5 секунд простоя





Split-Vote: Проблема

Долгая недоступность кластера на запись после потери лидера

Каждый раунд выборов, в результате которого не выбран лидер –
≥ 5 секунд простоя

В случае успешных выборов лидер появляется в самом начале раунда





Split-Vote: Проблема

Долгая недоступность кластера на запись после потери лидера

Каждый раунд выборов, в результате которого не выбран лидер –
≥ 5 секунд простоя

В случае успешных выборов лидер появляется в самом начале раунда

Если голоса разделились, оставшееся время раунда – простой





Split-Vote: Решение

У нас есть широковещательная рассылка отданных голосов





Split-Vote: Решение

У нас есть широковещательная рассылка отданных голосов

Каждый из узлов может вести подсчёт голосов за всех кандидатов





Split-Vote: Решение

У нас есть широковещательная рассылка отданных голосов

Каждый из узлов может вести подсчёт голосов за всех кандидатов

Стало понятно, что выборы никто не выиграет? Новый раунд





Split-Vote: Решение

У нас есть широковещательная рассылка отданных голосов

Каждый из узлов может вести подсчёт голосов за всех кандидатов

Стало понятно, что выборы никто не выиграет? Новый раунд

Экономия до 90% времени на каждый раунд с разделившимися голосами





Меню

- ✓ Raft overview
  - Термины: Journal, Term, LSN
  - Выборы
  - Гарантии
  - Ожидания != Реальность
- ✓ Raft / Tarantool: особенности
- Надстройки Raft
  - ✓ Pre-Vote
  - ✓ Split-Vote detection
  - **Fencing**





Fencing: Проблема

- Для записи в кластер необходимо обратиться к текущему лидеру





Fencing: Проблема

- Для записи в кластер необходимо обратиться к текущему лидеру
- Возможность существования нескольких нод, считающих себя лидерами одновременно (в разных term)





Fencing: Проблема №1

- Для записи в кластер необходимо обратиться к текущему лидеру
- Возможность существования нескольких нод, считающих себя лидерами одновременно (в разных term)
- Не более одного из “лидеров” сможет на самом деле произвести синхронную запись





Fencing: Пример №1

I’m leader

Term 15

1

Leader is 1

3

Term 15

Leader is 1

2

Term 15





Fencing: Пример №1

I’m leader

Leader is 1

Term 15

1

3

Term 15

Leader is 1

2

Term 15





Fencing: Пример №1

I’m leader

1

Term 15

I’m leader

3

Term 16

Leader is 3

2

Term 16





Fencing: Пример №1

I wonder why my writes timeout…

I’m leader

Term 15

1

I’m leader

3

Term 16

Leader is 3

2

Term 16





Fencing: Проблема №2

- При наличии pre-vote можно получить “заблокированный” кластер





Fencing: Пример №2

1

2

3

4

5





Fencing: Пример №2

3

1

5

2

4





Fencing: Пример №2

3

1

5

2

4





Fencing: Пример №2

3

1

5

2

4





Fencing: Пример №2

3

1

5

2

4





Fencing: Способы решения

- Для определения текущего лидера запрашивать у всех нод кластера текущий терм и состояние ноды (лидер или нет)
- Переложить определение лидера на внешнюю систему, гарантирующую его единственность, и довериться ей
- Доработать RAFT в Tarantool





Fencing: Способы решения

1. **Для определения текущего лидера запрашивать у всех нод кластера текущий терм и состояние ноды (лидер или нет)**

2. Переложить определение лидера на внешнюю систему, гарантирующую его единственность, и довериться ей
3. Доработать RAFT в Tarantool





Fencing: Способы решения

1. **Для определения текущего лидера запрашивать у всех нод кластера текущий терм и состояние ноды (лидер или нет)**
   - Не решает проблему заблокированного кластера

2. Переложить определение лидера на внешнюю систему, гарантирующую его единственность, и довериться ей
3. Доработать RAFT в Tarantool





Fencing: Способы решения

1. **Для определения текущего лидера запрашивать у всех нод кластера текущий терм и состояние ноды (лидер или нет)**
   - Не решает проблему заблокированного кластера
   - Увеличивается время недоступности на запись

2. Переложить определение лидера на внешнюю систему, гарантирующую его единственность, и довериться ей
3. Доработать RAFT в Tarantool





Fencing: Способы решения

1. **Для определения текущего лидера запрашивать у всех нод кластера текущий терм и состояние ноды (лидер или нет)**
   - Не решает проблему заблокированного кластера
   - Увеличивается время недоступности на запись
   - Требуется реализация данного решения во всех клиентах
2. Переложить определение лидера на внешнюю систему, гарантирующую его единственность, и довериться ей
3. Доработать RAFT в Tarantool





Fencing: Способы решения

1. Для определения текущего лидера запрашивать у всех нод кластера текущий терм и состояние ноды (лидер или нет)
2. **Переложить определение лидера на внешнюю систему, гарантирующую его единственность, и довериться ей**

3. Доработать RAFT в Tarantool





Fencing: Способы решения

1. Для определения текущего лидера запрашивать у всех нод кластера текущий терм и состояние ноды (лидер или нет)
2. **Переложить определение лидера на внешнюю систему, гарантирующую его единственность, и довериться ей**
   - Доверяем решение проблемы третьей системе
3. Доработать RAFT в Tarantool





Fencing: Способы решения

1. Для определения текущего лидера запрашивать у всех нод кластера текущий терм и состояние ноды (лидер или нет)
2. **Переложить определение лидера на внешнюю систему, гарантирующую его единственность, и довериться ей**
   - Доверяем решение проблемы третьей системе
   - Еще одна точка отказа
3. Доработать RAFT в Tarantool





Fencing: Способы решения

1. Для определения текущего лидера запрашивать у всех нод кластера текущий терм и состояние ноды (лидер или нет)

- Переложить определение лидера на внешнюю систему, гарантирующую его единственность, и довериться ей
- **Доработать RAFT в Tarantool**





Fencing: Как хочется

I'll look somewhere else then

No leader

Term 15

1

I'm leader

3

Term 16

Leader is 3

2

Term 16





Как хочется

3

1

5

2

4

I’m not leader

Ok, let’s vote!





Fencing: Решение

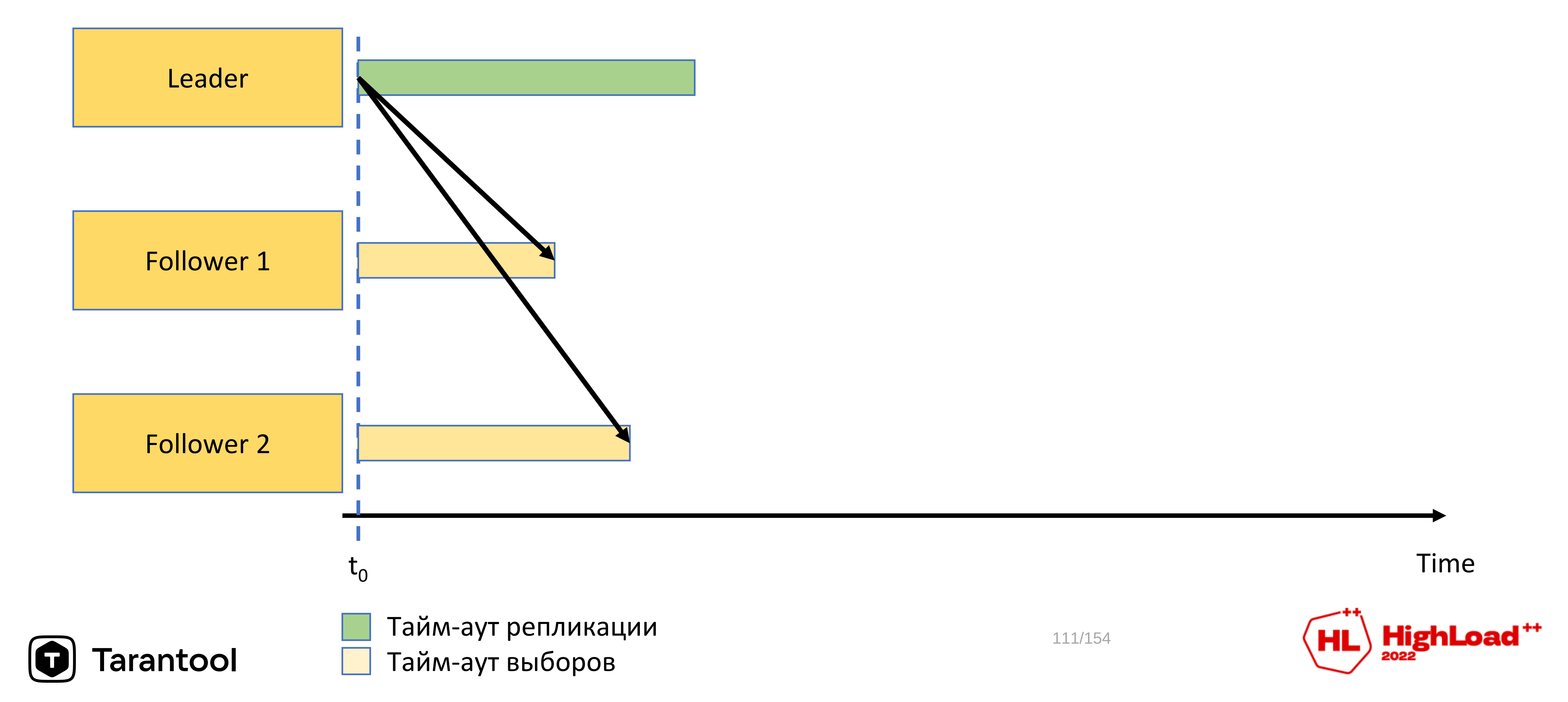
Нода, являющаяся лидером, следит за обрывом соединений к/от реплик

Если количество живых соединений стало меньше, чем текущий кворум, лидер снимает с себя полномочия

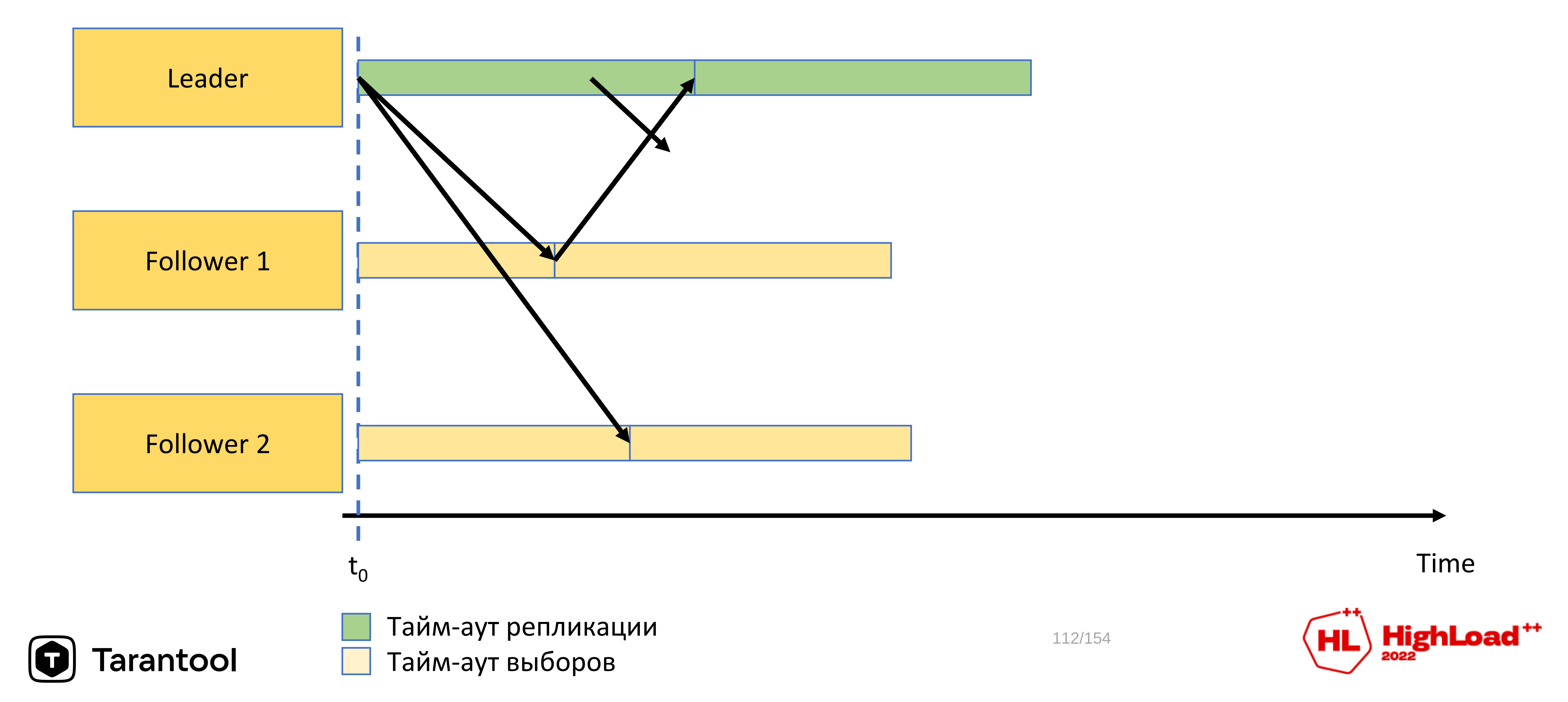




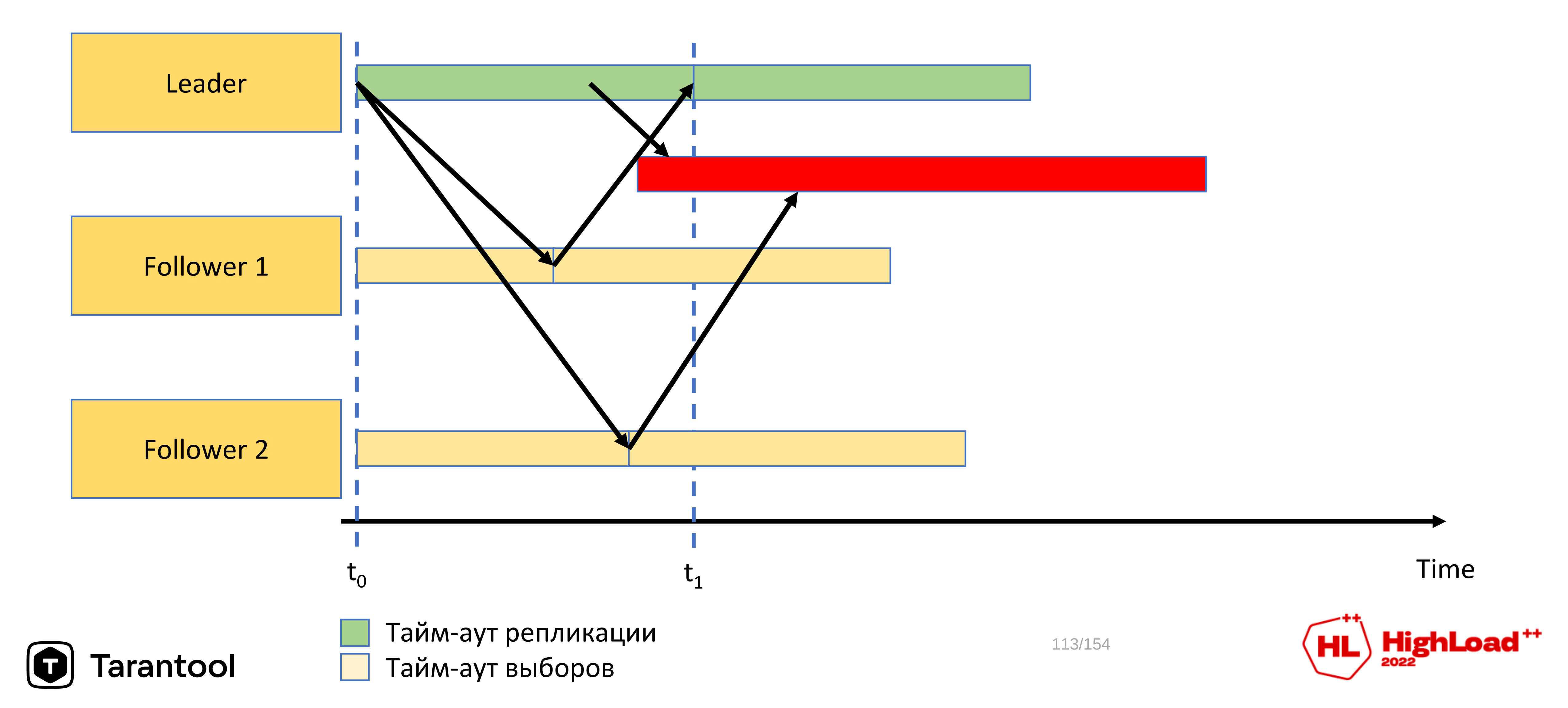
Fencing: Решение



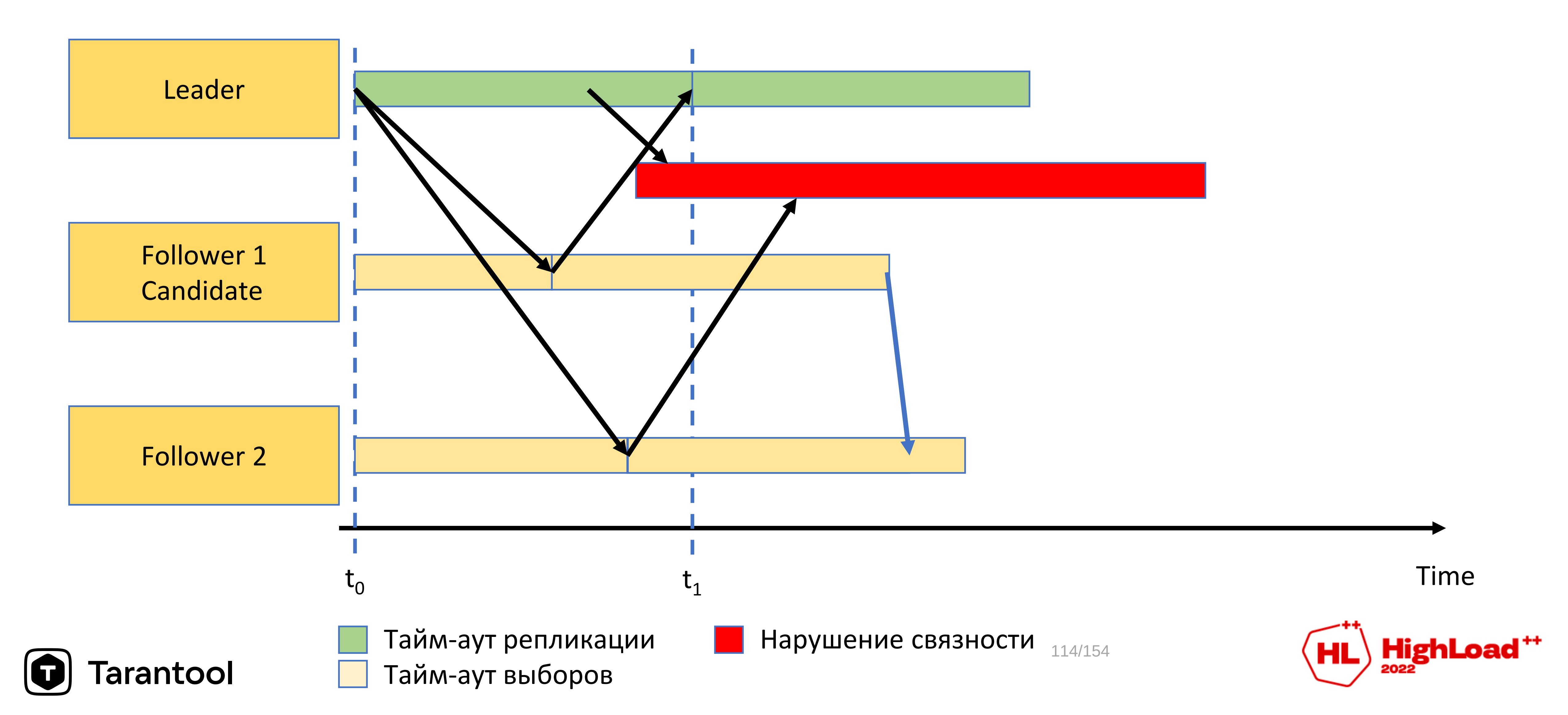
Fencing: Решение



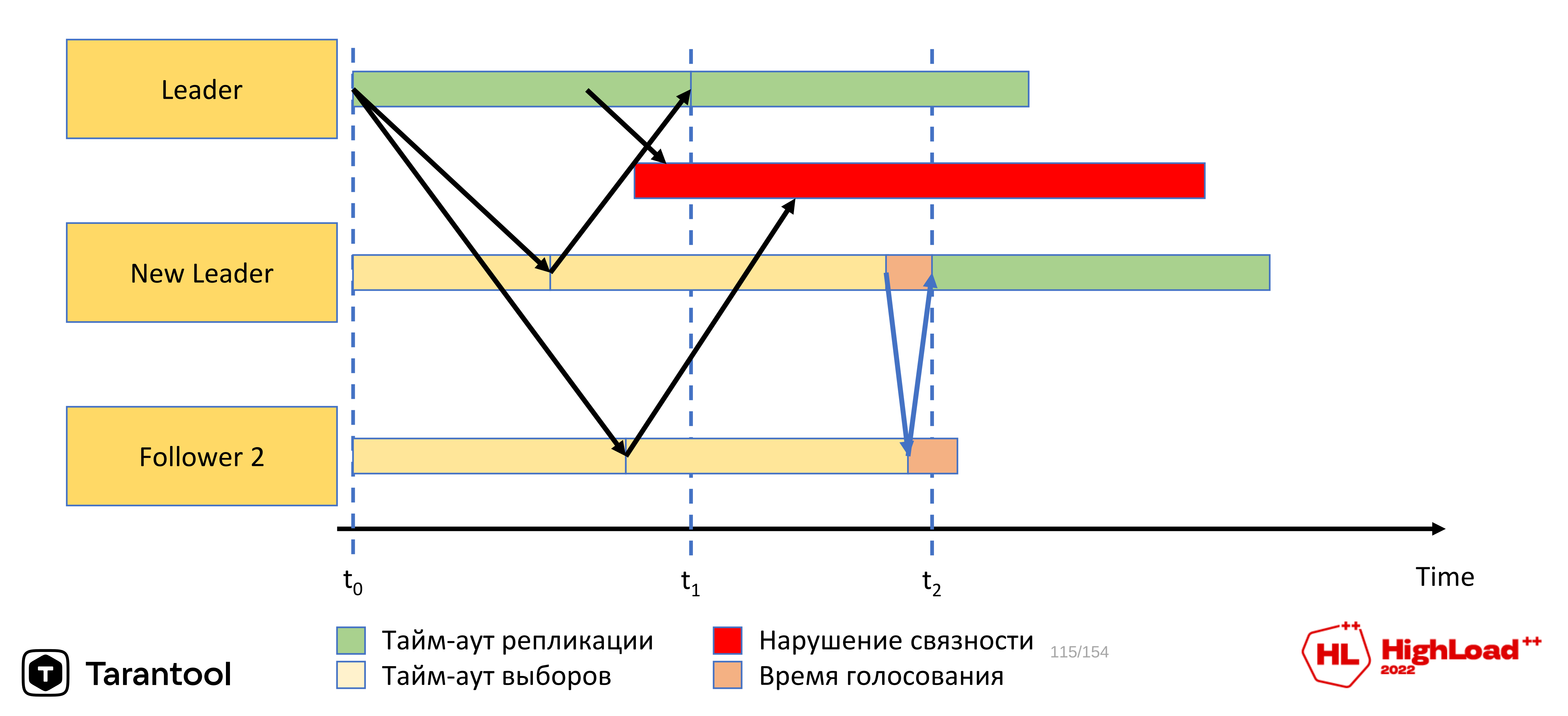
Fencing: Решение



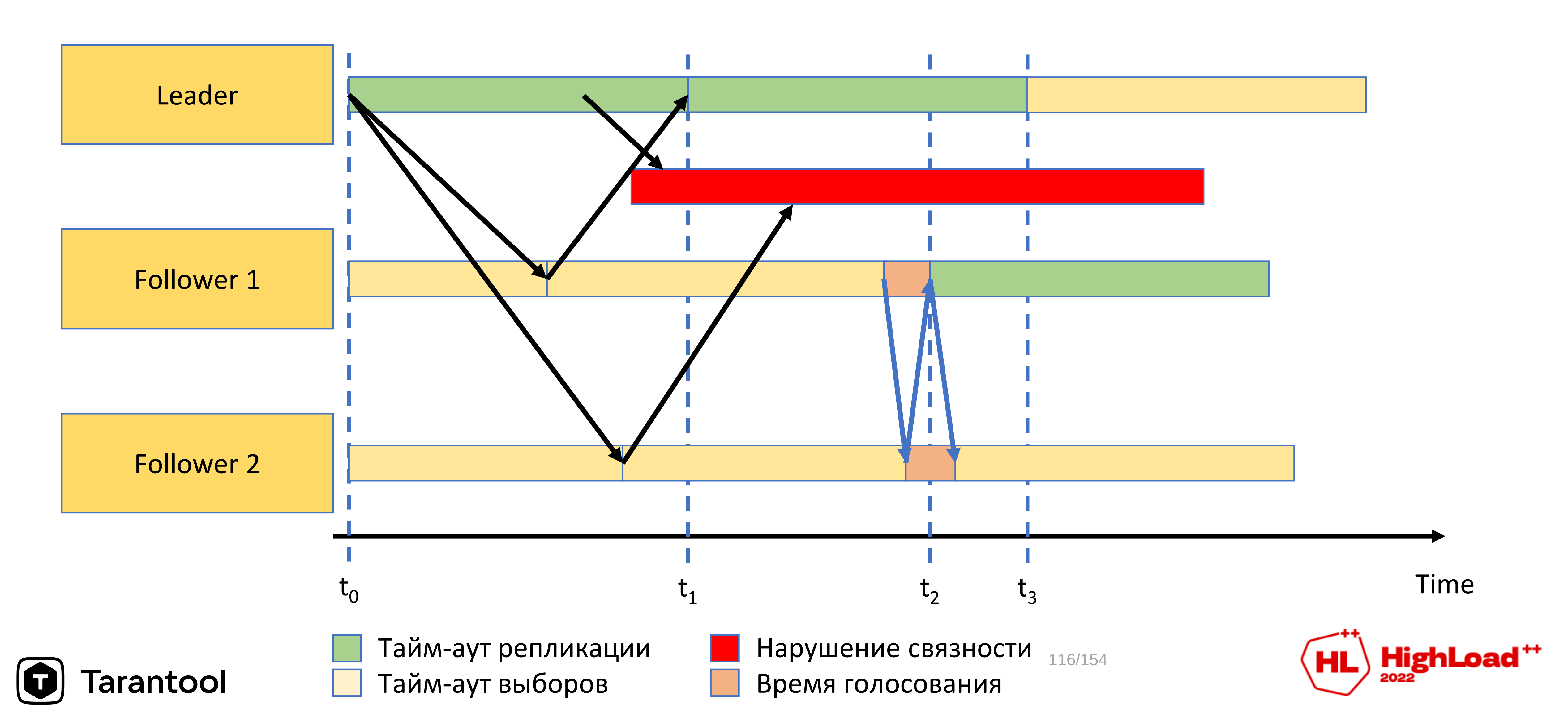
Fencing: Решение



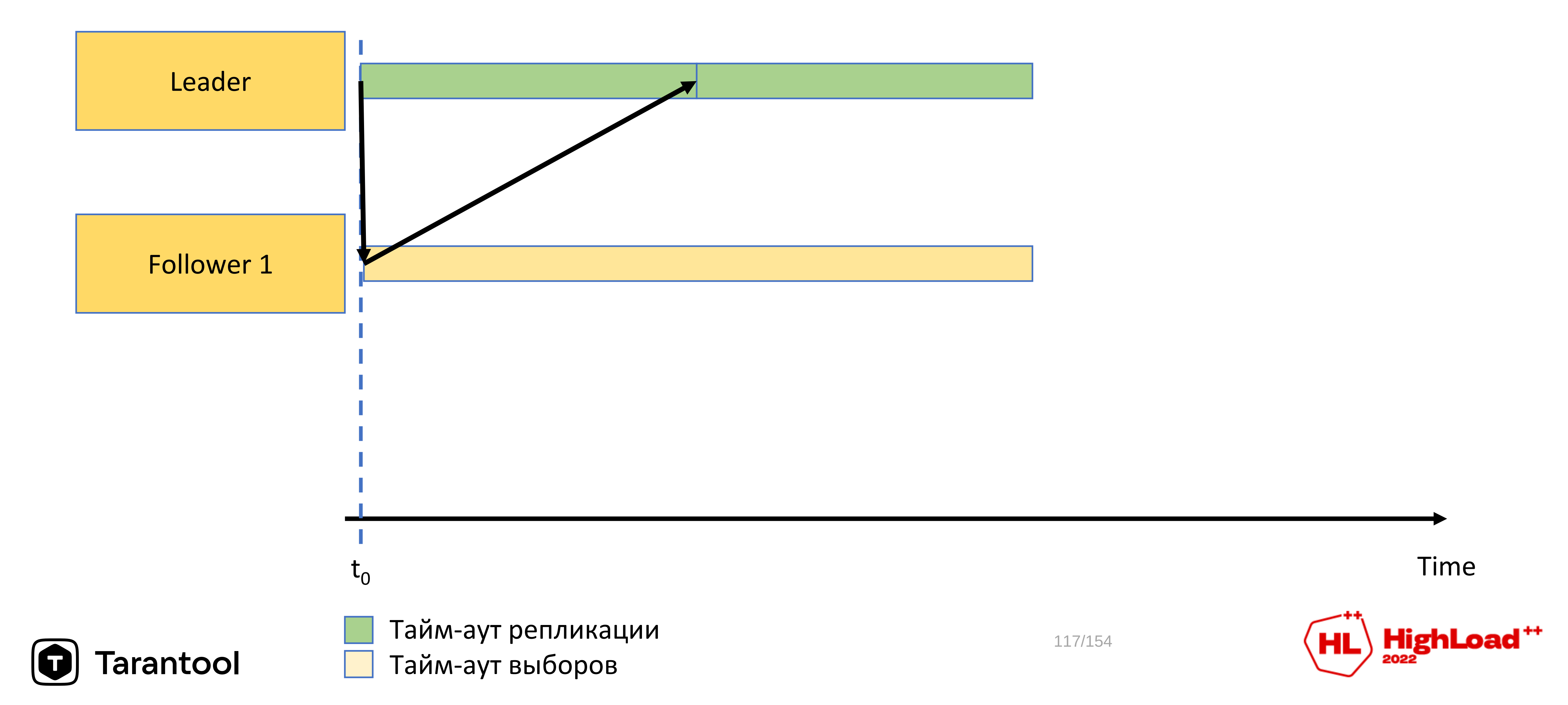
Fencing: Решение



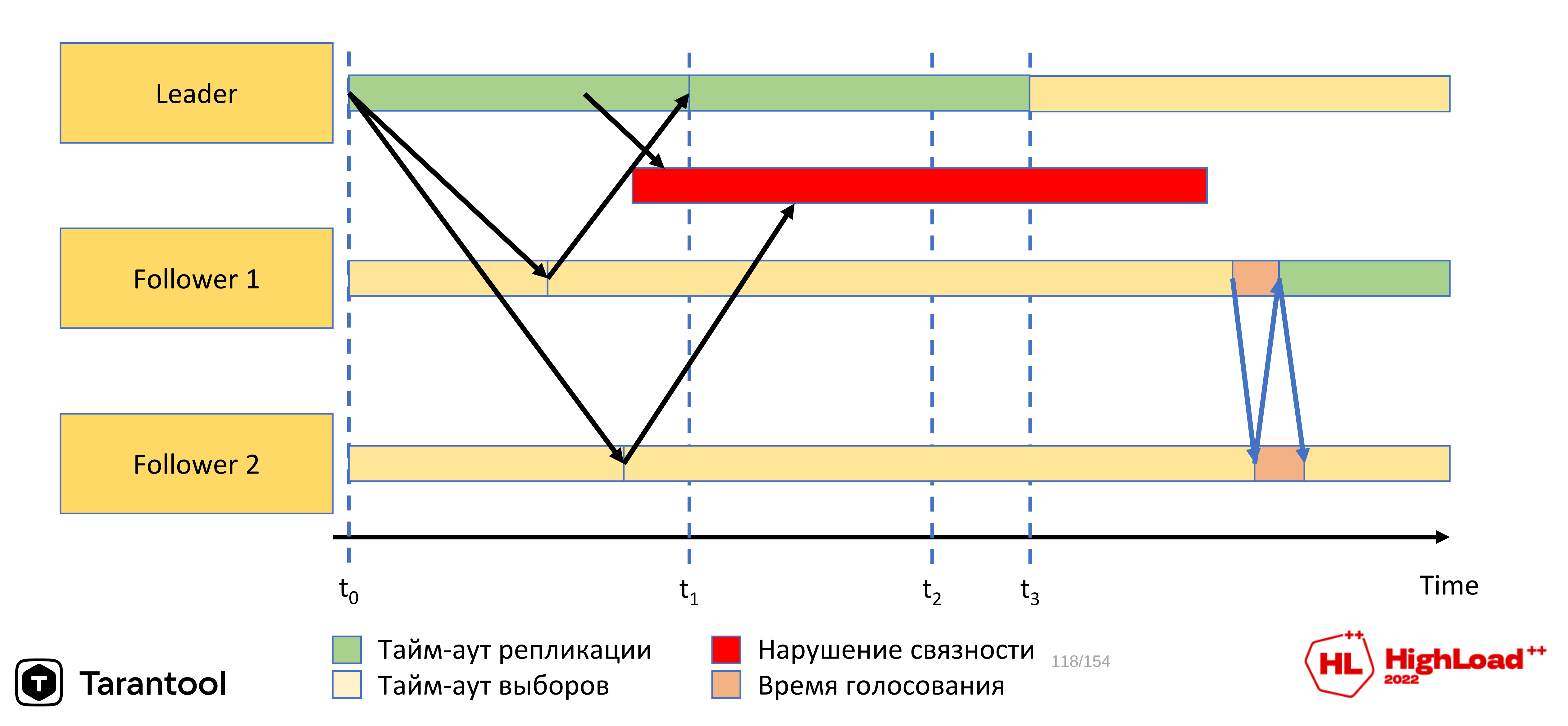
Fencing: Решение



Fencing: Решение



Fencing: Решение



Fencing: Решение

Три варианта применения Fencing:

1. Fencing выключен
2. Слабый fencing – более быстрый фейловер, но можно увидеть двух лидеров
3. Сильный fencing – точно не будет двух лидеров





Ожидали от Raft

- Один лидер на терм == один лидер в данный момент времени
- Лидер постоянен
- Быстрая смена лидера





Получили

- Один лидер на терм == один лидер в данный момент времени – Fencing
- Лидер постоянен – Pre-Vote, Fencing
- Смены лидера – Split-Vote detection позволил ускорить выборы ~10 раз





Обратная связь
и комментарии по докладу
по ссылке







Tarantool